the protein, so even the protein is not found in nature. The molecules (polypeptides) referenced in these claims are not found naturally but are contained as a component of a single large polypeptide (IMP dehydrogenase or IMPDH). Furthermore, reconstitution of the individual polypeptide fragments would not produce a system with the catalytic properties of IMPDH.

The purity of the polypeptides(s) is not a relevant issue because the coordinates of claims 5-7 were obtained from a crystal of *S. pyogenes* IMPDH. Such crystals are not found naturally but must be generated using complex screening matrices.

Claims 5, 6 and 7 are amended to relate "A crystalline molecule . . ." and there is no such product in nature, consequently the present invention is novel and non-obvious.

IV.   Those of Skill in the Art Would Understand What is Meant by "Binding Pocket"

The examiner admits that the specification "discloses" and defines "binding pocket," but still complains that the specification does not enable those of skill in the art "to predictably identify the binding pocket or the binding pockets from homologous sequences that are 60% or more identical." (Action, pages 3-4).

The invention is enabled in the ability to design inhibitors specific for bacterial forms of IMPDH. The target region for the design of such agents is the active site region of the molecule. This region is identified in the specification on page 10, line 15 via the statement; "protrusions on the proximal face of the a/b barrel scaffold range in size from 3-67 residues and define the character of the active site." The inventors specifically identified the amino acid residues comprising the active site region. These specific residues are listed in claims 5 and 7.

The basis for the involvement of these protrusions (loops) in defining the catalytic character of the bacterial enzyme is justified throughout the specification. Most notably, this information is contained on page 13-15 of the specification. This section describes structural features (active site helices H, J, and I) and specific residues involved in or contributing to catalysis. Further justification for the involvement of specific residues in catalysis is contained on pages 16-17 of the specification. This section describes the site-specific mutagenesis experiments in the active site region.

As is clear from the specification, the term "binding pocket" is equivalent to "catalytic pocket," "catalytic site," and "binding site." The Wilson patent cited by the examiner indicates this term is known to those of skill in the art. The description of the catalytic site and references

contained in the sepcificaiton describe the interaction of carious ligands (*e.g.* IMP and NAD) with IMPDH enzymes. For example all IMPDH enzymes have a cysteine at the active site and a consensus sequence for IMP binding (page 2 of the specification and reference cited herein). In view of this information it is reasonable to assume that the loaction of the binding pocket is generally conserved in all IMPDH enzymes thus fac characterized. Furthermore, IMPDH crystal structures indicate these enzymes all contain the same basic protein fold. The conservation of amino acid in sequence for bacterial IMPDH enzymes (Specification, Table 2), provides a method for the identification of amino acids corresponding to those residues found in the binding pocket of *S. pyogenes* IMPDH.

V.   Claim 6 is Not Too Broad

The examiner wants to limit the claims to the crystal specifically disclosed in the application. The examiner questions whether less than 100% sequence identity predictably gives a sequence with the same biochemical activity.

The examiner correctly asserts that in many cases it is impossible to predict the specific functional consequences of amino acid replacements. However, this is not globally true (*e.g.* it can be reliably predicted that replacement of the active site cysteine will have a substantial effect on function. The inventors describe consistent functional and sequence differences between bacterial and mammalian enzymes (specification page 2, lines 5-20; page 3, lines 2-11; page 5-8;and references, tables, and figures contained therein). The biochemical and kinetic differences between bacterial and mammalian enzymes are a consequence of the variance of specific, amino acid residues. Identification of these residues or combination of residues that impart this mammalian or bacterial enzyme signature is a prerequisite for the rational identification of agents that specifically target the bacterial enzyme. The invention enables the targeting of specific amino acids residues via site-directed mutagenesis or other means to elucidate the basis for the differences between bacterial and mammalian IMPDH enzymes. The percent homology definitions utilized in claim 6 represent the sequence diversity differential between bacterial and mammalian IMPDH enzymes.

VI.   Format of Claims 1-8 is Amended Without Altering the Scope of the Claims

Amendments are made to the format of claims 1-8. These amendments do not alter the scope of the claims. Amendments in a separately filed document are additions of SEQ IDs from

the concurrently filed computer sequence listing. Amendments herein spell out some of the terms. For example, the abbreviation "IMPDH" which is spelled out in the specification as inosine monophosphate dehydrogenase, and which only has one meaning, is spelled out in amended claim 1. The abbreviation term "IMP" is spelled out in the specification and only has one meaning (inosine monophosphate). Claim 8 is amended to also spell this term out. Because claims are interpreted in light of the specification, all abbreviated terms were definite to begin with. Amendments have not altered the scope of the claims.

Claim 4 adds on a step, which just repeats the preamble, so it does not further limit the claims.

A variety of molecular docking programs are available (two are provided herein as Exhibits A and B) to enable those skilled in the art to screen compounds to identify potential inhibitors. These programs are a common component of most structural biology software (e.g. Insight II from Molecular Simulations Inc.). Most of the newer programs provide for automated docking of ligands to receptors in the structure-based drug design process. A variety of purine analogues and mammalian inhibitors have been reported in the literature and would be likely candidates to begin a search for specific inhibitors of bacterial IMPDH enzymes. However, universities (e.g. University of Georgia) and commercial entities (Vertex Pharmaceuticals) have, and are continuing to construct, small molecule digital libraries. The present invention provides the structural coordinates of bacterial IMPDH that enables the screening of these candidate compounds/digital libraries.

Claims 5, 6, and 8 relate "binding pocket" which is clear.

"a computer generated representation of a molecule or molecular complex wherein said molecule or molecular complex comprises a binding pocket defined by the structural coordinates of *S. pyogenes* IMPDH amino acids ...."

SEQ ID numbers are filed concurrently so the amino acid numbering of claim 7 is clear. The amino acid numbers of claim 7 refer to Table 7.

Claim 7 is amended to insert the definition of "coordinates" from the specification. Those of skill in the art would understand the term without the amendment, and the amendment does not alter the scope of the claim because it merely recites the definition already in the specification.

Claim 9 is added.

VII. Zhang *et al.*, 1999, Is Prior Work of the Present Invention Published Less Than a Year Before Present Filing Date Therefore is Not Anticipated

Claims 1-8 were rejected under 35 U.S.C. § 102 (a) over Zhang *et al.*

Claims 1-8 are not anticipated by Zhang *et al.* A Declaration under 37 C.F.R. 1.131 from Dr. Collart is appended herein (Exhibit C) showing that the present invention was prior work of the present inventor published less than 1 year before the filing date of the present application. Moreover, as the examiner admits, not all elements of claims 1-8 of the present invention were taught by Zhang *et al.* Therefore, for both reasons Zhang *et al.* does not anticipate claims 1-8 of the present invention.

Disclosure in a publication does not prove that any "invention" within the meaning of 35 U.S.C. 102(g) has been made. An inventor's own prior work will not anticipate his later invention, absent a statutory bar. *In re Katz* 215 USPQ 14 (CCPA 1982), "a printed publication cannot stand as a reference under 102(a) unless it is describing the work of another." (*Katz* 215 USPQ 14, 17). Authorship of a publication does not raise the presumption of inventorship with respect to subject matter disclosed in publication (*Katz* 215 USPQ at 18). Therefore, co-authors may not be presumed to be co-inventors merely from co-authorship.

VIII. Sintchak Does Not Anticipate Claims 1-6 and 8

Claims 1-6 and 8 were rejected under 35 U.S.C. § 102(b). The examiner admits that

> however, it should be noted that patentable weight is given to the IMPDH composition, since the distinguishing features of the composition are in the specific sequence , structure, and function of IMPDH and not the cloning vehicle from which it came...It should be noted that the diffraction patterns obtained by Sintchak *et al.* would be useful for defining molecular structures of similar bacterial IMPDHs through the use of such methods as molecular structures of similar bacterial IMPDHs through the use of such methods as molecular replacement. The molecule and molecular complex solved by Sintchak *et al.* reads on a molecule comprising a homologue of the *S. pyogenes* IMPDH of the instant invention.

(Action, page 7)

Sintchak does not anticipate claims 1-6 and 8 because the bacterial enzymes and mammalian enzymes are different as described in the specification and Section IV herein. The structure of the Chinese hamster enzyme does not elucidate the basis for these differences. Only a crystal structure of a bacterial IMPDH enzyme can be used to delineate the basis for these differences.

The *S. pyogenes* and Chinese hamster IMPDH enzymes show only a 35% identity. This level of homology did not enable interpretation of the *S. pyogenes* model. A similar case was observed for the structure of IMPDH from *Tritrichomonas foetu*s which shows a similar level of identity to the *S. pyogenes* and Chinese hamster enzymes.

It should be noted that conditions for the generation of crystal in the present invention are different than those described for the Chinese hamster enzyme.

IX.   Collart *et al.* Does Not Anticipate Claims 5-6

Claims 5-6 were rejected under 35 U.S.C. § 102(b) over Collart.

Collart *et al.* does not anticipate claims 5 and 6 because Collart (1996) does not teach elements that are the same as claims 5 and 6. Collart (1996) describes amino acid similarities between IMPDH enzymes but does not teach methods to obtain well diffracting crystals of IMPDH enzymes, as in the present invention. Although IMPDH from *S. pyogenes* and *P. furiosus* share 50% sequence identity, crystals of *P. furiosus* IMPDH suitable for structure determination have not been obtained in any laboratory. There are methods based on amino acid sequence alignments to predict proteins that will crystallize. Every crystallization laboratory utilized a matrix (48-1536 individual conditions) of solutions to screen for conditions that produce well diffracting crystals. Despite the availability of these screens, many proteins still do not produce well diffracting crystals suitable for structure determination.

X.   Wilson Does Not Anticipate Claims 1-6 and 8

Claims 1-6 and 8 were rejected under 35 U.S.C. § 102(e) over Wilson *et al.*

> It should be noted that diffraction patterns obtained by Wilson *et al.* would be useful for defining molecular structures of bacterial IMPDH through the use of such methods as molecular replacement. The molecule and molecular complex solved by Wilson *et al.* reads on a molecule comprising a homologue of the *S. pyogenes* IMPDH of the instant invention.

(Action, page 8)

Wilson relates IMPDH Type II. Bacteria do not have this IMPDH, they only have Type I. There is only 35% identity between the proteins which made it impossible to use the crystal model to solve the bacterial IMPDH molecule structures.

XI.   Computer Sequence Listing

A computer sequence listing in accordance with 37 C.F.R. § 1.821-1.825 is filed concurrently.

XII.   Other Issues

The examiner did not consider the Holbrook publication in the Information Disclosure Statement because the date was not on the forms submitted.

The missing date for Holbrook in the Information Disclosure Statement is 1975. A copy of the publications was submitted, and the date was cited in the specification, therefore Holbrook should have been considered because the date of the publications was easily accessible in the record.

XIII.   Summary and Conclusion

For the reasons stated above, applicant requests allowance of all pending claims.

Please contact applicants' representative if you have any questions.

No other fees are believed due at this time, however, please charge any deficiencies or credit any overpayments to deposit account number 10-0435 with reference to our attorney docket number (21416-90042).

Respectfully submitted,
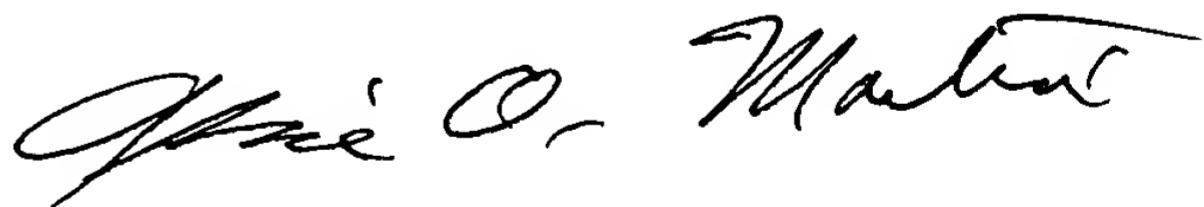
Alice O. Martin
Registration No. 35,601

BARNES & THORNBURG
2600 Chase Plaza
10 South LaSalle Street
Chicago, IL 60603
(312) 214-8316
March 14, 2001

CHDS01 AOM 92715v1

Critical Evaluation of Search Algorithms for Automated Molecular Docking and Database Screening

TODD J. A. EWING, IRWIN D. KUNTZ

Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California 94143-0446

ABSTRACT: The DOCK program explores possible orientations of a molecule within a macromolecular active site by superimposing atoms onto precomputed site points. Here we compare a number of different search methods, including an exhaustive matching algorithm based on a single docking graph. We evaluate the performance of each method by screening a small database of molecules to a variety of macromolecular targets. By varying the amount of sampling, we can monitor the time convergence of scores and rankings. We not only show that the site point–directed search is tenfold faster than a random search, but that the single graph matching algorithm boosts the speed of database screening up to 60-fold. The new algorithm, in fact, outperforms the bipartite graph matching algorithm currently used in DOCK. The results indicate that a critical issue for rapid database screening is the extent to which a search method biases run time toward the highest-ranking molecules. The single docking graph matching algorithm will be incorporated into DOCK version 4.0.   © 1997 by John Wiley & Sons, Inc.   *J Comput Chem* 18: 1175–1189, 1997

Keywords: automated molecular docking; 3D database search; molecular recognition; structure-based drug design

Introduction

With the advent of high resolution X-ray crystallography and NMR, structural chemists and biologists can study biomacromolec-

ular interactions in atomic detail. This information, combined with computational and visualization tools, has helped spawn the field of structure-based ligand design. A common step in the design cycle is the process of molecular docking, in which possible binding geometries of a molecule with a macromolecule are studied.

The docking process can be divided into two parts: a search algorithm and a scoring algorithm. The search algorithm should sample the degrees of freedom of the ligand:macromolecule system suffi-

ciently to include the true binding mode(s). The scoring algorithm should represent the thermodynamics of interaction sufficiently to distinguish the true binding mode(s) from all others explored.

Because of the computationally expensive nature of the search problem, many different solutions have been proposed. Docking with molecular dynamics and Monte Carlo algorithms has been explored,[1,2] including simulated annealing[3-5] and MCSS[6] methods. Other docking protocols consider molecular flexibility, including rotamer search,[7-9] distance geometry,[10] and genetic algorithm[11-13] methods. To make the search tractable for processing a large set of molecules, the molecular components are often treated as rigid objects. With this approximation, researchers have used systematic searching,[14] pattern recognition,[15-17] graph theoretical,[18-21] and other superposition[22] techniques to dock molecules.

The UCSF DOCK program belongs to the group of methods employing the rigid body assumption and uses graph theoretical techniques. Because of its speed, the program is often used to screen a large database of molecules, selecting potential ligands of a receptor target.[23] In this article, the term "ligand" is used loosely; it refers to any small molecule whose binding is under study. The term "receptor" refers to the macromolecule whose binding pocket is being explored.

The core of the DOCK search algorithm is the superimposition of ligand atoms onto predefined site points[24,25] that map out the negative image of the binding site. A matching process is used to determine which ligand atoms and site points are to be superimposed.[26] Multiple orientations are generated this way, with each receiving a score assessing the intermolecular interactions. This score is based on the intermolecular terms of a molecular mechanics force field.[27] Recently, an optimization procedure has been added that adjusts each orientation to improve the intermolecular interactions.[28,29]

In this work, we critically evaluate several matching algorithms for the docking process, including an exhaustive matching algorithm. The exhaustive algorithm was presented by Bron and Kerbosch as a method to detect cliques in an undirected graph.[30] It was later incorporated into a docking algorithm by Crippen and coworkers.[10,19] It has many attractive features, so we chose to evaluate it in the context of rigid molecular docking, score optimization, and database screening. As an exhaustive search, it avoids some of the artifacts encountered by the current matching

method. For example, with a nonexhaustive algorithm, adjusting parameters to increase the total amount of sampling can reduce the amount of sampling of certain binding modes.[28] Increased sampling, with the new algorithm, will always retain binding orientations found with less sampling, leading to a proper superset of binding modes. An exhaustive algorithm also does not require the additional parameters controlling the heuristics of the nonexhaustive search.[26]

Although the matching algorithms formally treat the ligand and receptor as rigid objects, they can readily be incorporated into a flexible docking scheme.[10,21,31-33] In future work, we will investigate how best to divide up a flexible docking problem into smaller rigid parts.

To evaluate the performance of the new matching algorithm, we propose a new assessment protocol based on screening a small database of molecules. Because we specifically intend to minimize any artifacts due to the quality of the scoring function in this work, we will not use experimental measurements as the standard, but instead, the global minimum of our current scoring function. We will also evaluate random, and partially random, search algorithms as controls with which to put the current DOCK performance in perspective. These control algorithms let us investigate fundamental issues of orientational sampling, such as the effect of using site points to guide the search.

Methods

BIPARTITE DOCKING GRAPH

Since the first release of DOCK, the search process has been driven by a matching procedure in which subsets of ligand atoms and receptor site points are identified that have equivalent internal distances.[18] Matching is formulated as a graph theoretical problem in which the ligand atoms and receptor site points are separate sets of nodes in a bipartite graph.[26] A match is defined as a set of compatible edges which connect a subset of ligand nodes with an equal number of receptor nodes. For the edges to be compatible, the distances among ligand nodes must map to equivalent distances among receptor nodes. An example of match formation is depicted in Figure 1. As this figure illustrates, to extend a match, all possible edges (including bad edges) must be considered; distance comparisons are used to identify and discard bad edges. Matches are extended until there is a

sufficient number of nodes in the match to define a unique orientation of the ligand.

Since version 2.0, DOCK avoided considering some bad edges with a pruning method involving distance binning.[26] Nodes were preorganized in distance bins, such that, for each seed node, sets of nodes in discrete distance intervals from the seed were identified. These bins guide match extension from a seed edge (connecting seed nodes), ensuring that candidate edges are compatible with the seed edge. They do not, however, ensure compatibility with other nonseed edges already included in the match. For example, the binning algorithm would avoid considering the bad edge in step II of Figure 1, but not the bad edge in step III. The storage requirements for this algorithm grow as $N_n N_b N_{n/b} \le N_n^3$, where N_n is the number of nodes (ligand atoms or receptor site points). N_b is the number of distance bins and grows with the longest distance and the inverse of the bin width. $N_{n/b}$ is the number of nodes in each distance bin which grows with N_n and the bin width.

SINGLE DOCKING GRAPH

Kuhl et al. proposed merging the bipartite docking graphs into a single docking graph,[19] which is then amenable to clique detection techniques developed by Bron and Kerbosch.[30] In a single docking graph, each node represents a pairing of an atom with a site point. Each edge identifies adjacent nodes, or two nodes for which both atom components and site point components are separated by equivalent distances. The docking graph is represented by an adjacency matrix in which each nonzero element identifies adjacent nodes. An adjacency matrix for the example depicted in Figure 1 is presented in Figure 2. The chief advantage with this representation is that all necessary distance comparisons are made during the construction of the adjacency matrix. Consequently, during matching, the adjacency matrix is used as a rapid filter to ensure that no bad edges are ever considered. This type of matching is presented in Figure 3. Although the single docking graph is one step removed from the intuitive appeal of the bipartite docking graph, it enables a more efficient solution to the docking problem.

The single docking graph representation has also been implemented in the FLOG docking program.[21] This program heavily prunes the matching search tree using a minimum-residual search heuristic. Although it examines all possible nodes at each branch point, it only pursues the node with
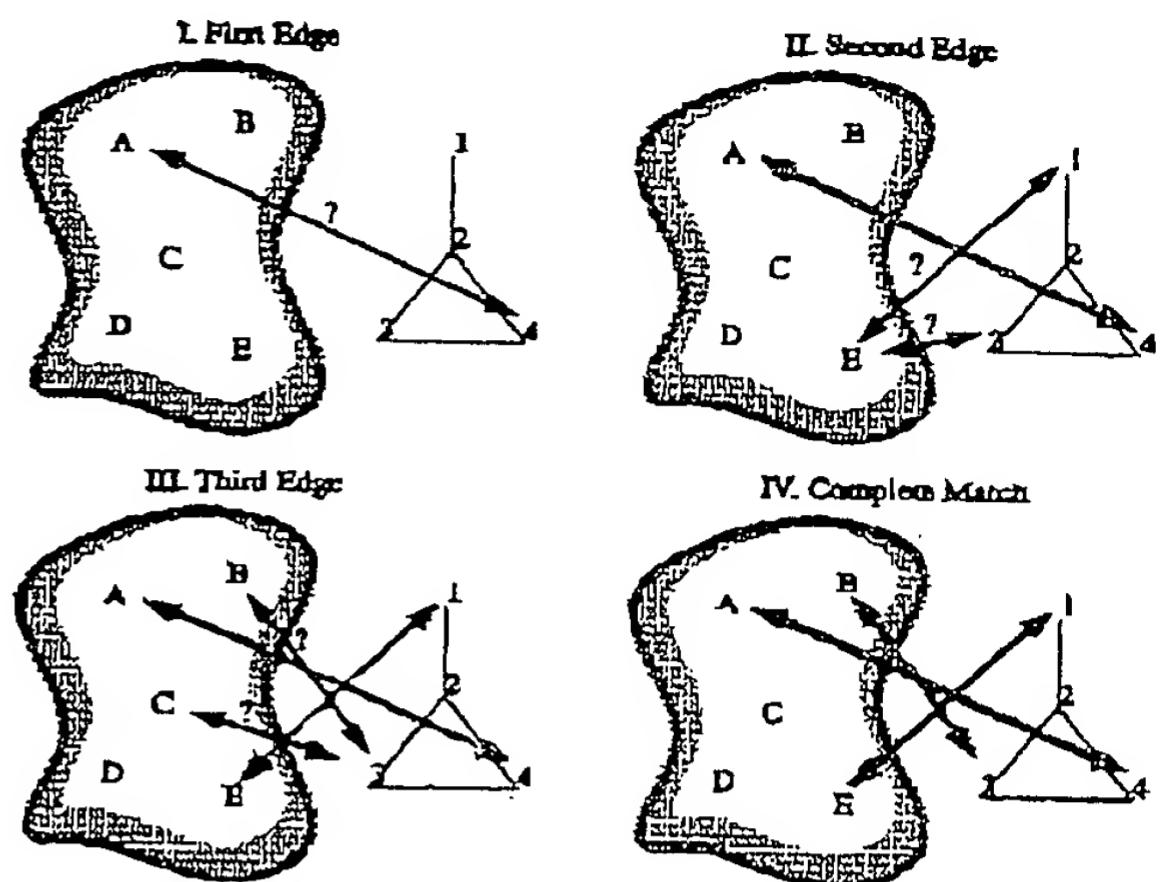

FIGURE 1. Bipartite graph matching algorithm. The receptor site points (A–E) and ligand atoms (1–4) are separate sets of nodes in a bipartite graph.[26] I. A first (seed) edge is considered. Of the 20 seed edges to be tried (5 points * 4 atoms), we first consider A4 for this example. In three dimensions, such a match would superimpose atom 4 onto site point A. This match would fix three of six orientational degrees of freedom. II. Second edges are considered. Of the 12 edges to be tried (4 points left * 3 atoms left), we consider E3. Because AE > 43, we discard E3 as a second edge. Then we consider E1. Because AE = 41, we retain E1 as a second edge. In three dimensions, this match would superimpose atoms 4 and 1 onto points A and E, respectively. This match would fix two more orientational degrees of freedom. III. Third edges are considered. Of the six edges to be tried (3 points left * 2 atoms left), we consider C3. Though AC = 43, EC < 13 so we must discard C3 as a third edge. Then we consider B3. Because AB = 43 and EB = 13, we retain B3 as a third edge. This match fixes the last of six orientational degrees of freedom. IV. The match is large enough to define a unique orientation which superimposes atoms 4, 1, and 3 onto site points A, E, and B, respectively.

the smallest difference in ligand and receptor distances with respect to the most recently added node in the match. Backtracking is only allowed at the seed level, where all possible nodes are pursued to initiate matching. As a result, the total number of matches can never exceed the number of nodes.

Here we propose implementing the single docking graph representation combined with a variation of the exhaustive clique detection method discussed by Bron and Kerbosch.[30] A clique is defined as a set of fully adjacent nodes (i.e., a completely connected subgraph) which cannot be
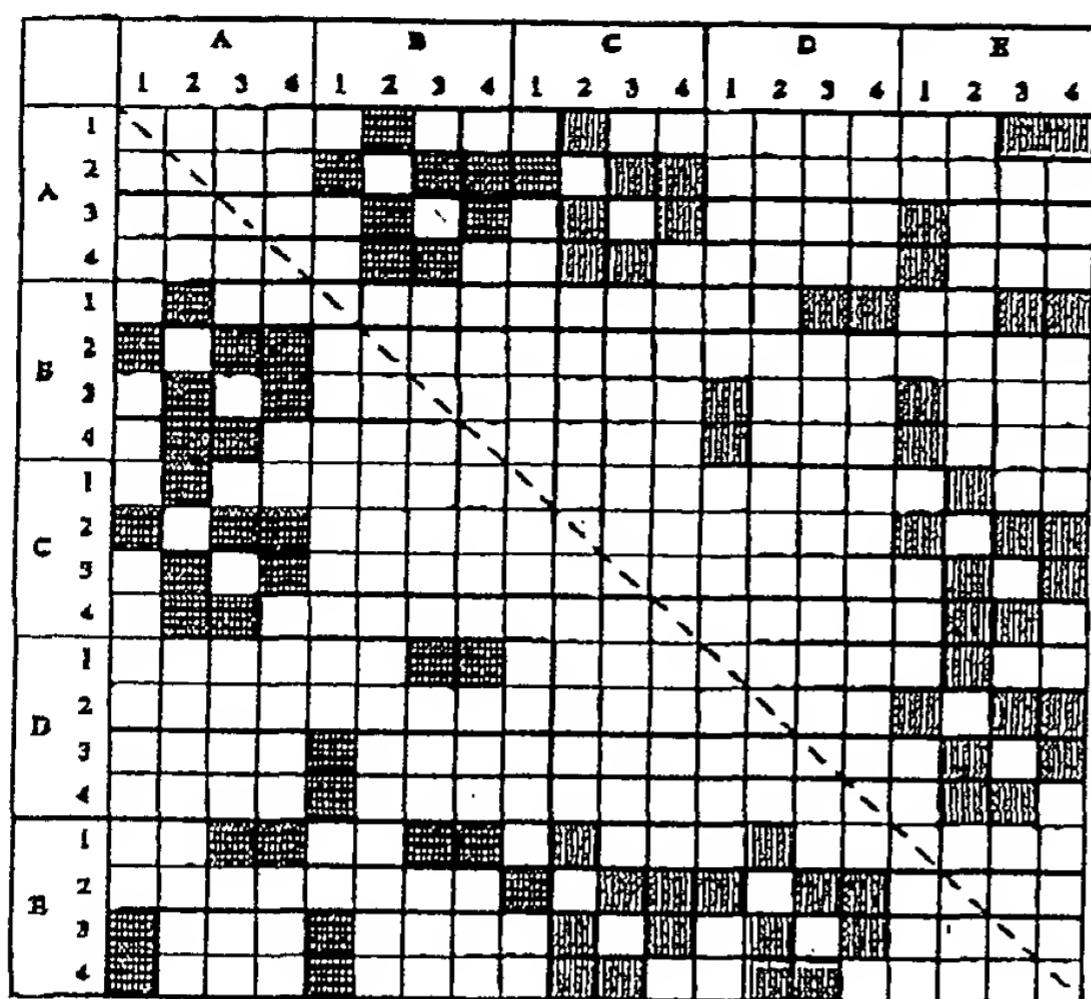
FIGURE 2. Adjacency matrix for single docking graph algorithm. This matrix identifies all adjacent nodes for the example given in Figure 1. Each node is defined as a site point–ligand atom pair, for instance, A4. For two nodes to be adjacent, the intra-atom distance must be equal to the intra-site point distance. For example, matrix element (A4, E1) is turned on because $\overline{AE} = \overline{41}$. The matrix is symmetric.

further enlarged without adding a nonadjacent node. Much attention is given to the intractable nature of the maximum clique problem. It is classified as NP-complete because the solution time grows faster than any polynomial expression of the problem size.[19] For the application of molecular docking, however, we are not trying to find the single, largest clique. The process of matching is in fact a process of finding completely connected subgraphs within an undirected graph: a less restrictive, and therefore more tractable problem than finding cliques and maximum cliques. Although Bron and Kerbosch actually present two methods and recommend a bounding technique for clique detection, we find their original, brute-force method sufficient for finding fully connected subgraphs in a manner efficient for molecular docking.

MATCHING PARAMETERS

In our molecular docking implementation, we use two parameters to determine node adjacency: a distance tolerance and a distance minimum. The distance tolerance parameter addresses experimental uncertainty in the ligand and target structures.
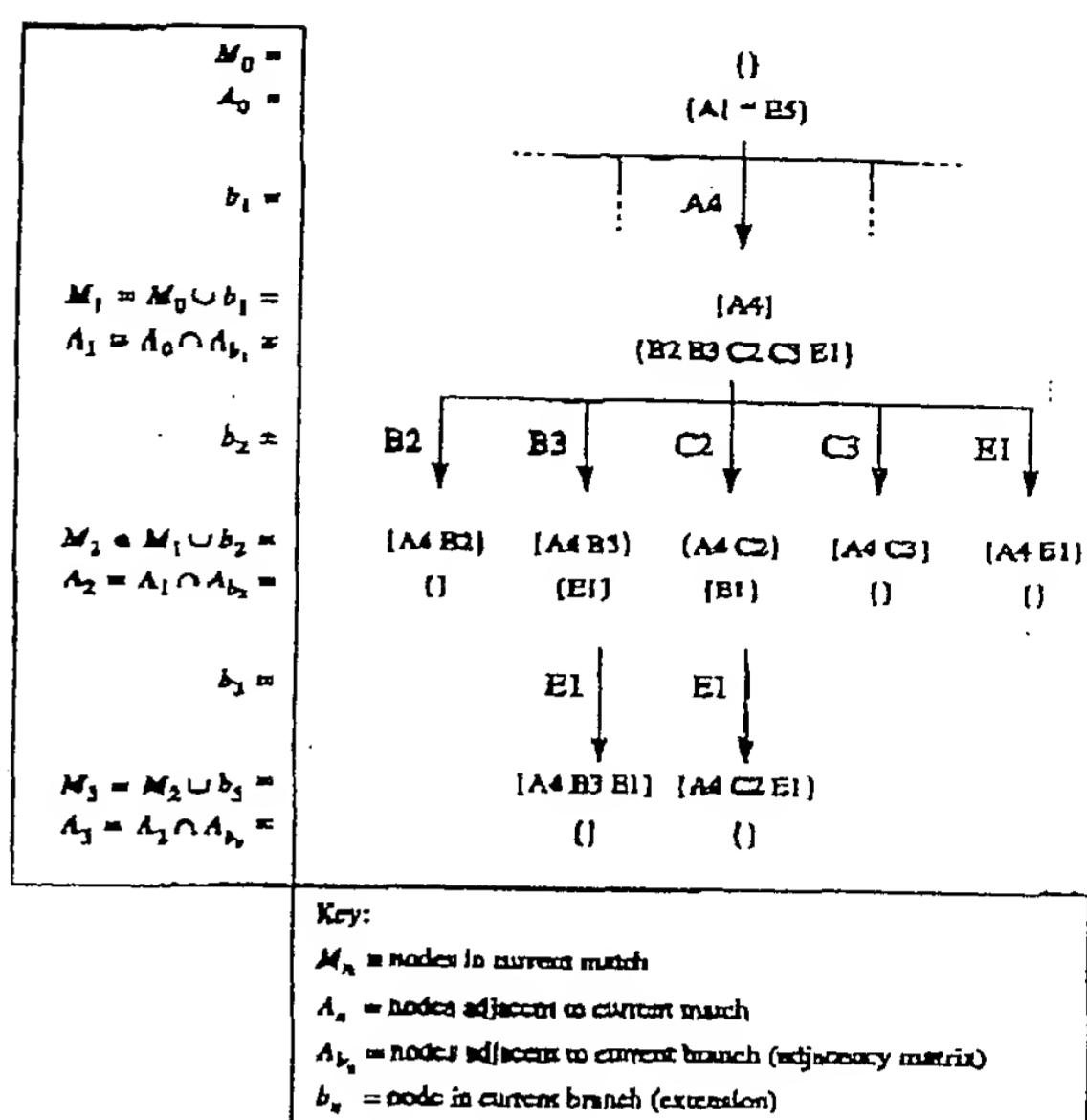

FIGURE 3. Single docking graph matching algorithm. This figure depicts the entire search path starting from the same seed used in Figure 1. Matching begins with the match set, M_0, empty and the adjacency set, A_0, maximally filled. A match is extended by finding the union of the previous match, M_{n-1}, and a branch node, b_n, which is selected from A_{n-1}. The new adjacency set, A_n, is the intersection of A_{n-1} with the set of all nodes adjacent to b_n, which is A_{b_n} and is taken from the adjacency matrix presented in Figure 2. Match extension continues until A_n is empty. Matches with three or more nodes define a unique ligand orientation (see text and Fig. 4). The match set, {A4 B3 E1}, corresponds to the solution presented in Figure 1.

The distance minimum parameter enables the search to focus on the longer, more relevant internal distances. If adjacency information is stored in a matrix, then, for most docking situations, the matrix can be very large. The matrix size grows as $(N_{lig} N_{rec})^2$, where N_{lig} is the number of nonhydrogen ligand atoms and N_{rec} is the number of receptor site points. Because these matrices are sparse (ca. 1% elements typically occupied), we store only the nonzero elements of each row of the matrix as an integer list of nodes. The probability, p_{on}, of an element being nonzero is a function of the distance tolerance and distance minimum. The memory requirement for the adjacency lists is $p_{on}(N_{lig} N_{rec})^2$. We presort each adjacency list so that the process of finding the common elements of two lists (the $A_n = A_{n-1} \cap A_{b_n}$ steps in Fig. 3)

can be performed on a once-through basis at a speed comparable to the use of the complete matrix.

The advantage of an exhaustive algorithm is that, when sampling is increased, the search is guaranteed to include the search space explored at the lower sampling level. This property helps to avoid sampling artifacts encountered with the bipartite matching algorithm.[28] Despite its exhaustive nature, it does not undergo a combinatorial explosion for larger systems because of user control over the sampling parameters. Typical sampling parameters for a docking scenario having fewer than 30 ligand nonhydrogen atoms and fewer than 50 target site points are: four nodes minimum for a match, and 0.5-Å distance tolerance and 2.0-Å distance minimum for node adjacency. Because the search never explores invalid branches, search time grows as a function of the number of distance-constrained solutions rather than the number of possible unconstrained solutions. Therefore, docking larger molecules into larger sites can be made nearly as rapid if a smaller distance tolerance or larger distance minimum is chosen. Because memory reserved for the adjacency lists is dynamically allocated, the memory burden is adjusted as well.

MINIMUM MATCH SIZE

Some confusion exists in the literature over how many atoms and site points must be in a match to define a unique orientation. The orientation is generated by a ligand transformation which has a translation and a rotation component. The translation vector has three degrees of freedom. The rotation matrix has four degrees of freedom. Three of them are represented by the Euler angles. The fourth is represented by the sign of the determinant. A rotation with a positive determinant retains the handedness of the object it transforms, while one with a negative determinant will reverse the handedness of the object. If one knows in advance whether to reverse the handedness of the object, then only the three Euler angles need to be determined. For example, when docking a chiral ligand in which only one stereoisomer is relevant (e.g., protein or peptide ligands), only the positive-determinant rotation matrix would be of interest. When docking a ligand available as a racemate, then both transformations would be of interest. The FLOG program,[21] for instance, routinely samples both mirror images a ligand, even when the ligand is achiral. When the sign of the determinant is known in advance, the six degrees of freedom of
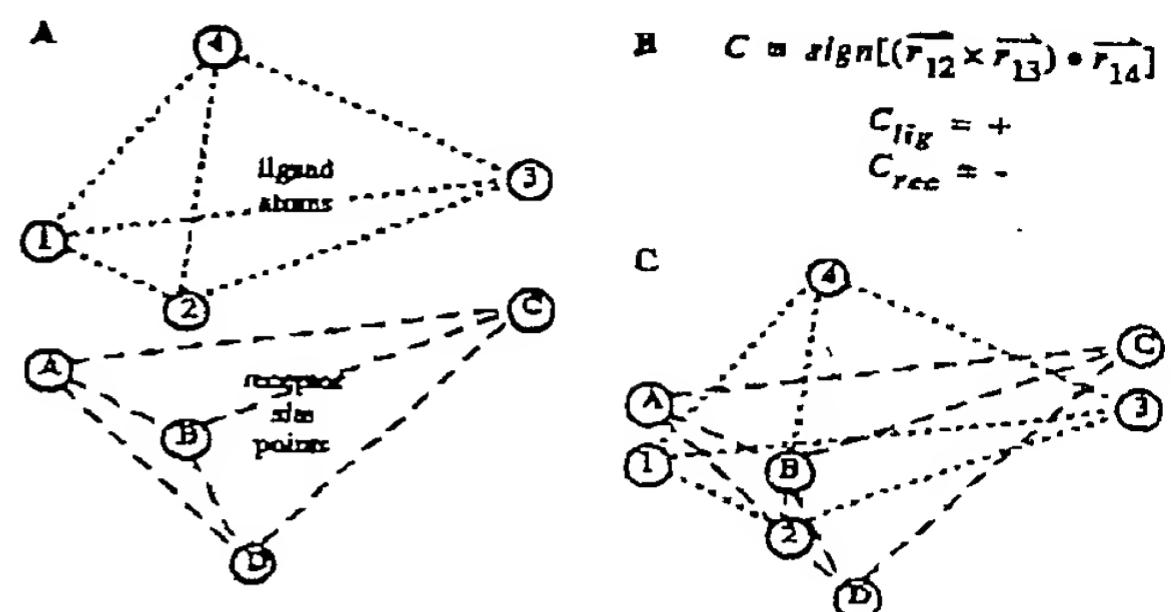

FIGURE 4. Relative chirality of match points. (A) Distance matching might identify a set of atoms {1 2 3 4} to match with a set of site points {A B C D}, such that A is with 1, B is with 2, and so on. Although the internal distances within the atoms are equivalent to those within the site points, the handedness is opposite. (B) We define the relative chirality, C, according to the sign of the triple product. For any given four-point match, the probability that the relative chiralities are the same is 50%. (C) If the chirally opposed sets are superimposed without inverting the chirality of the ligand set, then the resulting least squares fit is poor.

the rotation and translation are uniquely determined by a match set containing three nonlinear atoms and site points.

Processing a larger match set causes only one transformation to be allowed. As illustrated in Figure 4, when a one-to-one mapping has been made between two sets of four nonplanar points, then each set can be assigned a relative chirality. This is true even if the points come from an achiral molecule or from a set of site points where chirality is ambiguous. If the relative chiralities are the same, then the ligand can be oriented normally. If the relative chiralities are opposite, then either the ligand is inverted when oriented, or if that is not desired, then the match is discarded. In fact, all larger matches that are supersets of the discarded match are also discarded, because they too yield inconsistent matches. If these steps are not taken, then the resulting orientations will poorly superimpose the ligand atoms and receptor site points in the match set, even though all the distance tolerances are met (Fig. 4C).

ADDITIONAL SEARCH CONSTRAINTS

The systematic design of the matching algorithm makes it well suited to incorporate specialized search constraints. Some examples, although not assessed in this study, are mentioned because

they have been shown to be useful elsewhere. To avoid oversampling particular binding modes, orientational degeneracy checking has been studied.[17,29] In the new matching algorithm, a degenerate orientation is detected as a degenerate match whose nodes are a subset of nodes in a larger match. In other words, only subgraphs that are true cliques need be processed. As another example, chemical information can be used to guide the matching process using labeled atoms and site points. Only nodes composed of a chemically compatible atom and site point are used to seed or extend a match. Much like the repellent node implementation of Kuhl et al.,[19] matches adjacent to chemically incompatible nodes are discarded. In addition, sampling can be focused on particular regions of the active site by defining critical site point clusters. This technique is similar to the approach used in targeted DOCK[34] and FLOG,[21] except that clusters can be of arbitrary size and number. The matching process automatically restricts itself to make sure all matches include members from each cluster. The new matching

algorithm lends itself so well to these constraints that when activated, they contribute a negligible computational overhead, and can lead to considerable speed improvements for database searches.[21,34]

CONTROL METHODS

We will test a total of five methods (Table I) to isolate specific aspects of the search process. These methods range in complexity from a completely random search to the bipartite and single graph procedures described previously. We begin with the uniform random transformation (URT) method, which explores a predefined rectangular volume enclosing the active site. It is the most simple and "hypothesis-free" of the methods tested here. URT will indicate the minimum level of performance that we expect from any docking algorithm. The uniform random matching (URM) method explores the irregularly shaped volume described by the collection of site points. It will show the performance gains, if any, of using a "negative image"

TABLE I.
Search Methods.

Abbrev.	Method	Description	Hypothesis Tested
URT	Uniform random transformation	• Construct rectangular volume enclosing site • Randomly move molecule center of mass within volume • Randomly rotate • Each molecule in database sampled uniformly	• Random method used as reference • Rectangular enclosure is sufficient Site point description is unnecessary
URM	Uniform random matching	• Match random subsets of atoms with random subsets of site points • Superimpose match atoms onto match site points with a least squares fit • Each molecule in database sampled uniformly	• An irregular volume to describe the site is more efficient
BRM	Biased random matching	• Match randomly (like URM) • SGM controls amount of sampling for each molecule in database	• An irregular volume is more efficient • Spending more time on molecules which match better is more efficient
SGM	Single graph matching	• Using single graph, exhaustively match subsets of atoms with subsets site points with equivalent internal distances (DOCK 4.0)	• Fitting some atoms precisely onto of some site points is more efficient
BGM	Bipartite graph matching	• Using bipartite graph, nonexhausitively match using binning algorithm (DOCK 3.5)	• Fitting some atoms precisely onto some site points is more efficient • Binning algorithm more efficient

approach to map out the binding site. The biased random matching (BRM) method is identical to URM, except it uses the new matching algorithm to determine the number of random matches to try for each molecule. Once the number of matches has been determined, BRM uses completely random selections of nodes to form the actual matches used to generate molecule orientations. Because BRM is a hybrid approach, it is meant to help isolate the source of any differences between URM and the new matching algorithm. The single graph matching (SGM) algorithm uses the new matching algorithm to determine both the number of matches and the actual orientations to try for each molecule. SGM will reveal the performance gains, if any, of using site points to not only map out the most interesting binding site volume, but to also direct the positioning of individual ligand atoms. The bipartite graph matching (BGM) method is the existing DOCK 3.5 matching algorithm. It will reveal the advantage, if any, of using a nonexhaustive search method with a longest distance first heuristic. BGM is described last because, within the spectrum of different search methods, its algorithm is the most elaborate.

TEST SYSTEM

We assess the performance of the search methods in the following way. We dock a set of 100 molecules, chosen randomly from the set of uncharged, medium-sized, and generally rigid molecules in the available chemicals database (ACD).[35] In our study, a medium-sized molecule is one with 15 to 35 nonhydrogen atoms. A generally rigid molecule is one with no single bonds except those attaching hydrogen atoms, attaching terminal nonhydrogens (i.e., methyl or hydroxyl groups), or participating in ring structures. Molecules meeting these three criteria compose 40% of the ACD. For each molecule in the test set, a single CONCORD[36]-generated conformation is used.

We see several advantages to using such a data set of molecules to test the search methods. First, the docking conditions represent a close approximation to the typical application of DOCK to database screening. Not only can we study the convergence of score for each molecule, we can study the convergence of relative scores, or rankings, of the set of molecules. Second, the docking conditions allow us to explore a multitude of diverse molecular shapes so that our results are less subject to potential artifacts of a particular ligand:receptor system. Although some may argue that

studying a set of known, potent ligands would be more relevant, we counter that the databases DOCK searches often do not contain potent binders, and that DOCK frequently finds micromolar inhibitors to serve as lead compounds.[23] By choosing a random subset of molecules, we, in fact, will arrive at a set that best represents the typical array of molecules tested. By biasing the subset to include medium-sized, generally rigid molecules, we also focus on that portion of the database which is best treated by rigid molecular docking.

For each search method, we perform multiple docking runs, and vary the amount of sampling from zero to a value at which the docking results converge. The key sampling parameters for each method are listed in Table II. Scoring and opti-

TABLE II.
Sampling Parameters.

URT	
Total orientations	0–900,000[a]
URM	
Total orientations	0–100,000[a]
Nodes min/max	3/3[b]
BRM	
Distance tolerance	0–0.9 Å[a]
Distance minimum	2.0 Å[c]
Nodes min/max	3/3[b]
SGM	
Distance tolerance	0–0.6 Å[a]
Distance minimum	2.0 Å[c]
Nodes min/max	4/10[d]
BGM	
Distance tolerance	1.5 Å[e]
Ligand bin width	0.1–0.9 Å[a,f]
Receptor bin width	0.1–0.9 Å[a,f]
Ligand bin overlap	0.1–0.5 Å[a,f]
Receptor bin overlap	0.1–0.5 Å[a,f]
Nodes min/max	4/4[e]

[a] For some test systems, the upper limit was not reached if docking results converged early.

[b] Set to three, because, as the size of a random match increases, the least-squares superposition procedure increasingly biases the orientation toward the centroid of the site points.

[c] Set large enough to exclude atoms sharing a covalent bond.

[d] Minimum of four chosen so that chirality could be used to filter matches. Maximum of ten is somewhat inconsistent with value chosen for BGM, but we presume any effects of this would be small.

[e] Chosen as historical default.

[f] Minimum value not zero because of a numerical instability of the algorithm.

mization parameters are listed in Table III. The range of values in these tables correspond to timings of less than 0.1 second/molecule to more than 100 seconds/molecule on a modern workstation (see subsequent text).

For each docking run, several properties are computed that compare the results from any specific run to the best results from all runs combined (assumed to contain the global minimum). These properties are summarized in Table IV. When these values are plotted versus time, the convergence of each property can be monitored. We assume that a better search algorithm will lead to more rapid convergence.

When considering the average behavior of each property, we compute both the usual mean and also a rank-weighted mean. The rank-weighted mean is more sensitive to the behavior of the top scoring molecules, which are of most interest in database screening runs. Although many kinds of weighting functions could be chosen for this pur-

TABLE III.
Scoring and Optimization Parameters.

Type	Force field
Bump maximum	3[a]
Dielectric	4r
Grid spacing	0.3 Å
Interpolation	Trilinear
Convergence criteria	0.1 kcal/mol[b]
Maximum iterations	500[c]

[a] Nonzero maximum allows some orientations with limited Van der Waals clashes with the receptor to be recovered by the minimizer.
[b] A relatively tight convergence criteria was selected to reduce noise in the score evaluation, so that differences between methods were more directly attributable to differences in sampling. The rank correlation would be especially vulnerable to such noise.
[c] A large iteration limit was also selected to reduce noise in the score evaluation by preventing the minimizer from terminating prematurely.

TABLE IV.
Comparison Methods.

Method	Definition	Equation	Rank weighted equation	Range
Average relative score	• For each molecule in the docking run, normalize its score by the best score it ever received in the site • Then compute the average over the molecules in the set	$$\dfrac{\sum_{i=1}^{N} S_i/S_i^{min}}{N}$$	$$\dfrac{\sum_{i=1}^{N} \dfrac{S_i/S_i^{min}}{i}}{\sum_{i=1}^{N} \dfrac{1}{i}}$$	[0, 1]; unitless
Rank Correlation	• Assign a rank, y_i, to each molecule based on its best score in the docking run • Then correlate y_i with the rank of each molecule based on the best score it ever received in the site, x_i	$$\dfrac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$	$$\dfrac{\sum_{i=1}^{N} \dfrac{1}{i}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} \dfrac{1}{i}(x_i - \bar{x})^2}$$	[-1, 1]; unitless
Average RMS error	• For each molecule in the docking run, compute the RMS error of its predicted orientation compared to that which received the best score for that site • Then, compute the average over the molecules in the set	$$\dfrac{\sum_{i=1}^{N} R_i}{N}$$	$$\dfrac{\sum_{i=1}^{N} \dfrac{R_i}{i}}{\sum_{i=1}^{N} \dfrac{1}{i}}$$	[0, ∞]; in angstroms

TABLE V.
Receptor Structures from PDB[37,38] Used for Test Systems.

Code	Structure	Resolution	R factor	Site description
121D	DNA dodecamer with Netrospin[39]	2.2 Å	0.198	Site is broad, presenting two continuous binding sites in the major and minor grooves of the DNA dodecamer; highly polar
1ULB	Purine nucleoside phosphorylase with guanine[40]	2.75 Å	0.204	Site has two pockets; one is broad and centrally located, another where actual ligand binds is peripheral and solvent excluded
3DFR	Dihydrofolate reductase with NADPH and methotrexate[41]	1.7 Å	0.152	Site has a deep, centrally located binding pocket; mixed polar and nonpolar regions
4FAB	Fab fragment with fluorescein[42]	2.7 Å	0.215	Site is shallow with three pockets formed by the six hypervariable loops; generally nonpolar
9HVP	HIV-1 protease with A-74704[43]	2.8 Å	0.182	Site is a long, narrow tube, which completely penetrates protein; mixed polar and nonpolar regions

pose, we chose to use the reciprocal of the rank for convenience.

To make sure that our conclusions are generalizable, we analyzed the methods using five different receptor sites listed in Table V. These sites were chosen from the list of complexes of high resolution, well-refined structure with a ligand having a well-defined binding position. They were also chosen based on very different-shaped binding sites. The chief features of each site are shown in Table V.

Results and Discussion

DOCKING CONDITIONS

The test cases were prepared for docking in the standard way. Site points were constructed using the sphere generation accessory program of DOCK with default parameters.[18] We selected the cluster of site points which occupied the binding site of the actual ligand in the crystal complex. Within this cluster, we merged the positions of tightly grouped site points using a 2-Å cutoff. The final number of site points used for each receptor ranged from 30 to 60.

RESOURCE USAGE

All docking calculations were performed on Silicon Graphics Indigo2 workstations, equipped with 200 MHz R4400 processors and 128 MB of RAM, so timings are consistent among the different methods. Several weeks of computer time were required to complete all runs. All methods required approximately 13 MB of RAM to store the scoring grids. The URT and URM methods required negligible additional memory for matching and orienting. BRM and SGM required up to 0.1 MB of RAM for matching arrays. BGM required 1 MB of RAM for matching arrays.

TEST SYSTEMS

Selected results for the 3DFR test system are presented in Figures 5, 6, and 7 to illustrate the type of data we collected. As shown in Figure 5, the weighted average score generally converges asymptotically to an optimum as sampling increases. The scores from all matching methods converge to within 90% of the optimum in about 10 seconds/molecule, whereas the URT method requires about 100 seconds/molecule. The weighted rank correlation in Figure 6 also shows convergent behavior, but with some interesting differences. It goes through much wider fluctuations, indicating that small changes in score have large effects on the rankings of the top scoring molecules. It appears to discriminate among the different methods, selecting BRM and SGM as superior, BGM and URM as next best, and URT as worst again. In particular, BRM and SGM both show a rapid initial rise, indicating that, with very
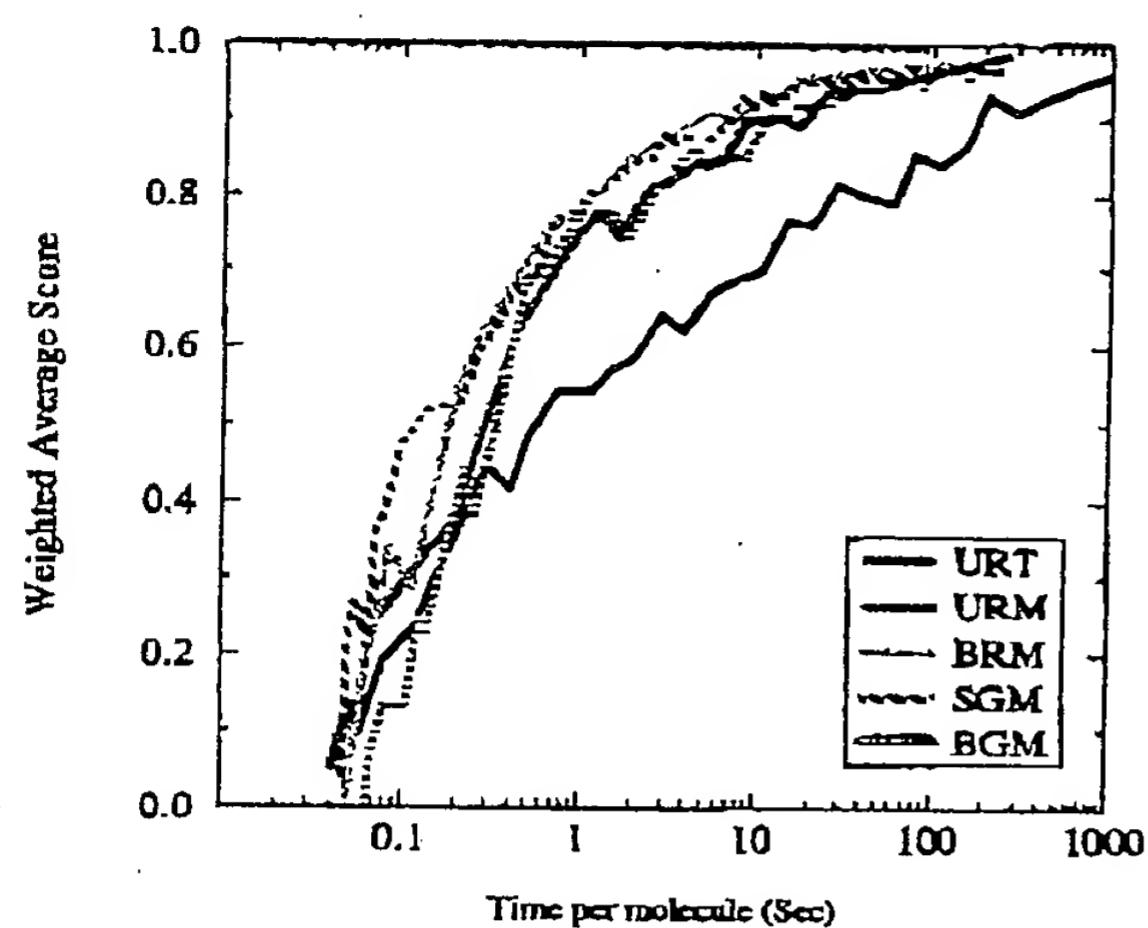
FIGURE 5. Weighted average score for molecules docked to 3DFR using SPHGEN site points. Each curve represents a search algorithm in Table I. Each data point is a weighted average of the score for all molecules in a particular run using the equation in Table IV.

little sampling, these methods come closest to predicting the rankings of the top scoring molecules. The convergence of weighted RMSD in Figure 7 indicates how long it takes the different methods to predict reproducibly the same binding mode of the top scoring molecules. Although the two top performing methods converge in predicted score


FIGURE 6. Weighted rank correlation for molecules docked to 3DFR using SPHGEN site points. Each data point is an average of the rank correlation for all molecules in a particular run using the equations in Table IV.

FIGURE 7. Weighted average RMSD for molecules docked to 3DFR using SPHGEN site points. Only the two best search algorithms from Table I are presented. Each data point is an average of the RMSD for all molecules in a particular run using the equations in Table IV.

and ranking in about 10 seconds, they require about 500 seconds before they consistently predict the same binding mode. This result indicates that, for these molecules in this site, several good scoring orientations must exist that are close in score but distant in space. We found similar results for the other sites as well.

We found the weighted and unweighted forms of the average score and rank correlation to be the most relevant in assessing database screening performance. This gives us four measurements of five methods over five sites. Instead of presenting 100 different curves, we have condensed each curve into a single value—the convergence time—which represents the time at which the 90% threshold value is passed (and not recrossed). In Table VI, we present the convergence times along with the speed improvement factor of each method compared to the URT method.

SCORE CONVERGENCE

With respect to the unweighted average score in Table VI(A), all matching methods show roughly equivalent convergence and outperform URT by a factor of 10. Therefore, on average, site points provide a much more succinct description of the active site than the smallest enclosing box, especially when searching a site that is large or difficult

TABLE VI.
Convergence Properties of Search Methods.[a]

Site	Search Methods				
	URT	URM	BRM	SGM	BGM
(A) Unweighted Average Score					
121D	200 (1x)	10 (20x)	20 (10x)	10 (20x)	8 (30x)
1ULB	200 (1x)	10 (20x)	20 (8x)	20 (10x)	8 (20x)
3DFR	100 (1x)	9 (10x)	10 (10x)	10 (10x)	20 (6x)
4FAB	30 (1x)	7 (4x)	8 (4x)	7 (5x)	7 (4x)
9HVP	200 (1x)	7 (30x)	5 (40x)	5 (40x)	6 (30x)
Mean	100 (1x)	9 (10x)	10 (10x)	9 (10x)	9 (10x)
(B) Rank-weighted average score					
121D	400 (1x)	10 (40x)	6 (60x)	3 (100x)	8 (50x)
1ULB	1000 (1x)	60 (20x)	30 (40x)	20 (70x)	30 (40x)
3DFR	200 (1x)	20 (9x)	10 (20x)	8 (30x)	20 (10x)
4FAB	90 (1x)	7 (10x)	4 (20x)	4 (20x)	7 (10x)
9HVP	200 (1x)	7 (30x)	2 (100x)	4 (40x)	5 (40x)
Mean	300 (1x)	20 (20x)	7 (40x)	6 (50x)	10 (30x)
(C) Unweighted rank correlation					
121D	400 (1x)	20 (20x)	10 (30x)	7 (60x)	10 (40x)
1ULB	1000 (1x)	60 (20x)	40 (20x)	30 (30x)	80 (10x)
3DFR	400 (1x)	50 (9x)	20 (20x)	10 (30x)	40 (10x)
4FAB	200 (1x)	30 (7x)	10 (20x)	20 (7x)	50 (3x)
9HVP	700 (1x)	20 (40x)	8 (90x)	10 (60x)	20 (50x)
Mean	500 (1x)	30 (10x)	20 (30x)	20 (30x)	30 (20x)
(D) Rank-weighted rank correlation					
121D	400 (1x)	20 (30x)	10 (40x)	7 (60x)	20 (30x)
1ULB	1000 (1x)	60 (20x)	20 (60x)	20 (70x)	80 (20x)
3DFR	600 (1x)	100 (6x)	10 (40x)	10 (40x)	20 (30x)
4FAB	400 (1x)	10 (30x)	3 (100x)	4 (90x)	10 (30x)
9HVP	300 (1x)	7 (40x)	2 (100x)	4 (80x)	3 (80x)
Mean	500 (1x)	20 (20x)	7 (70x)	8 (60x)	20 (30x)

[a] For each sampling method and receptor site, we find the amount of sampling, in seconds, beyond which all values are within 90% of the maximum. With the time of URT as the reference, the relative speed factor of each method is reported in parentheses. The geometric mean over all receptor site values is reported at bottom. Because of the large uncertainty in these values, all values are rounded to one significant digit.

to define *a priori*. Because the site points were generated based on general considerations of shape, this result should be generalizable to other "negative image" techniques, like the shape-based critical point methods of Nussinov and Wolfsen,[44] and the energetic probe methods of Goodford.[45] The rank-weighted average score in Table VI(B) shows more discrimination among the matching methods. While the two uniform random methods (URT and URM) had more difficulty converging with the top-scoring molecules, BRM and SGM actually converged more quickly. This implies that trying a uniform number of orientations for the

molecules in the database is inefficient with respect to processing the top-scoring molecules. BGM performed better than URM, but not as well as either BRM or SGM.

RANK CONVERGENCE

The convergence of the rank correlation further confirms the differences between methods. The unweighted rank correlations in Table VI(C) again show a general 10- to 30-fold advantage of using site points to dock the molecules. Interestingly, the time required to get the rank correct is two- to

fivefold greater than the time to get the average score [Table VI(A)] correct. The weighted rank correlations in Table VI(D) also further discriminate among the methods. URM and BGM still show the 20- to 30-fold advantage over URT. BRM and SGM again outperform URT by 60- to 70-fold. Of all measures, this last one is arguably the most relevant to database screening, because absolute scores are generally not used as strict cutoffs, but instead the rankings are used to select some subgroup of molecules whose number is amenable to further processing. Most often, it is the top-scoring subgroup of most interest, so using a rank-weighted correlation should focus our attention on how the methods treat this particular subgroup of molecules. Therefore, it appears that, of the methods investigated here, BRM and SGM are the best suited for database screening.

BGM VERSUS SGM

The single graph matching method clearly outperforms the existing bipartite graph matching method by up to twofold in speed. This result may appear counterintuitive, because SGM is exhaustive, whereas BGM uses heuristics to speed the search. However, the precomputing of the adjacency matrix and the rapid processing of the adjacency lists show that reformulating the problem into an efficient form can be just as effective as using heuristics. SGM has the additional advantage of requiring fewer fundamental matching parameters than BGM (Table II).

BRM VERSUS SGM

Why does a simplistic random matching algorithm, BRM, perform so competitively with the distance matching algorithm, SGM, and even outperform BGM? Does this suggest that distance matching is an unnecessarily complicated solution to docking? We seek to resolve this question by breaking the problem into three parts.

1. What is the disadvantage of sampling each molecule uniformly?

2. Why does distance matching sample molecules nonuniformly?

3. Is an orientation from random matching just as good as one from distance matching?

First, spending the same time on each molecule may result in spending too little time on the better

scoring molecules. One feature of the force field scoring function is that it tends to favor larger molecules.[21] For the set of molecules used in this study, we have plotted the best score for each molecule against its size in Figure 8. Although there is some trend, the correlation coefficient is not large. A stronger trend exists in Figure 9, relating molecule size to bump filtering. Large molecules have a greater propensity to bump into receptor atoms when oriented in the site. Because we use a bump filter in DOCK to discard poor orientations before the more computationally expensive scoring and optimization steps, we are more likely to discard an orientation of a large molecule than that of a small molecule. Forcing a uniform number of matches per molecule would then result in a size-biased attrition through the bump filter and, overall, spending less time on the potentially better scoring, larger molecules.

Second, nonuniform sampling arises in distance matching because the number of matches is related to the number of internal distances that the ligand has in common with the site points. Larger molecules have more internal distances, and so will tend to have more in common. Therefore, larger molecules tend to generate more matches than smaller molecules. However, a molecule that is not larger, but rather similar in shape to the binding site, will also have more internal distances
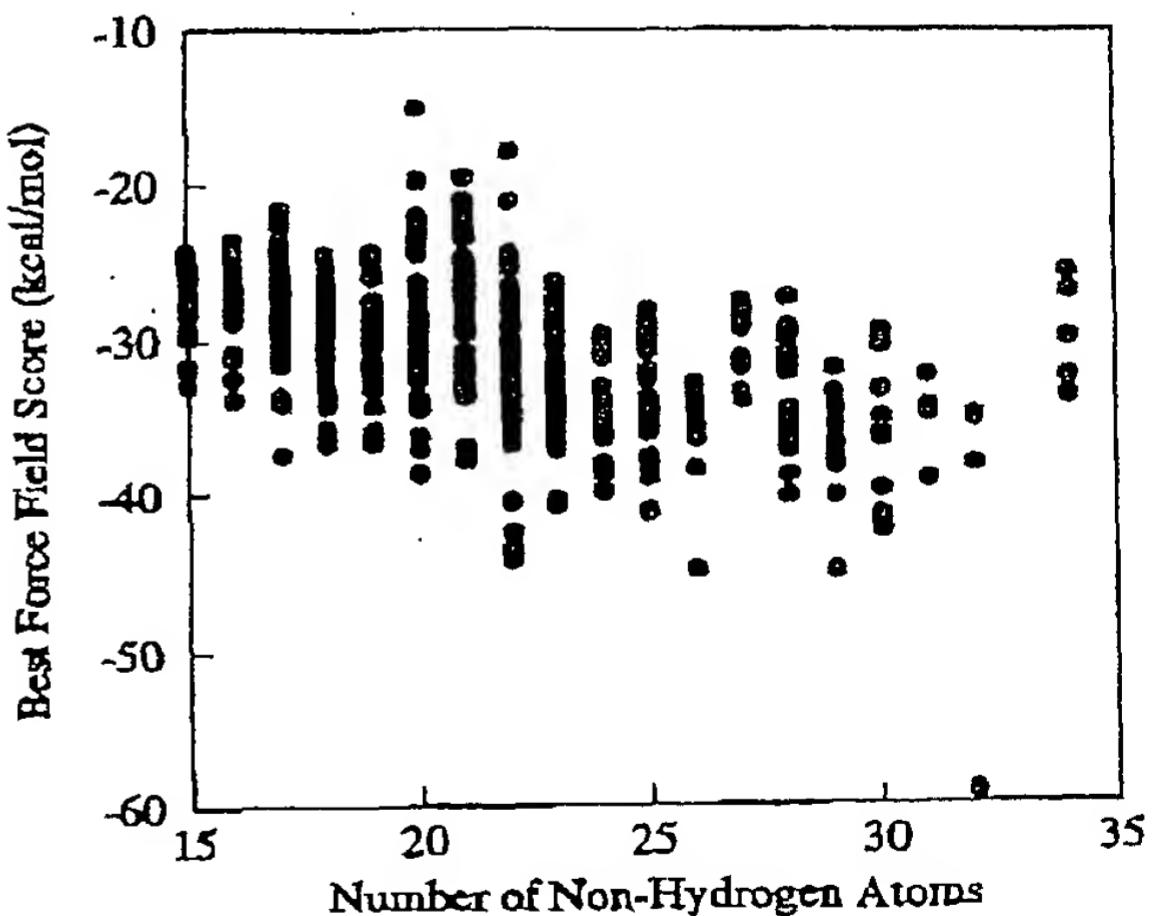

FIGURE 8. How molecule size affects force field score. The best force field score for each molecule is plotted versus the number of nonhydrogen atoms in the molecule. The plots for each receptor are pooled into this single plot to show the overall trend. Fitting a line to these data yields an R^2 value of 0.24.
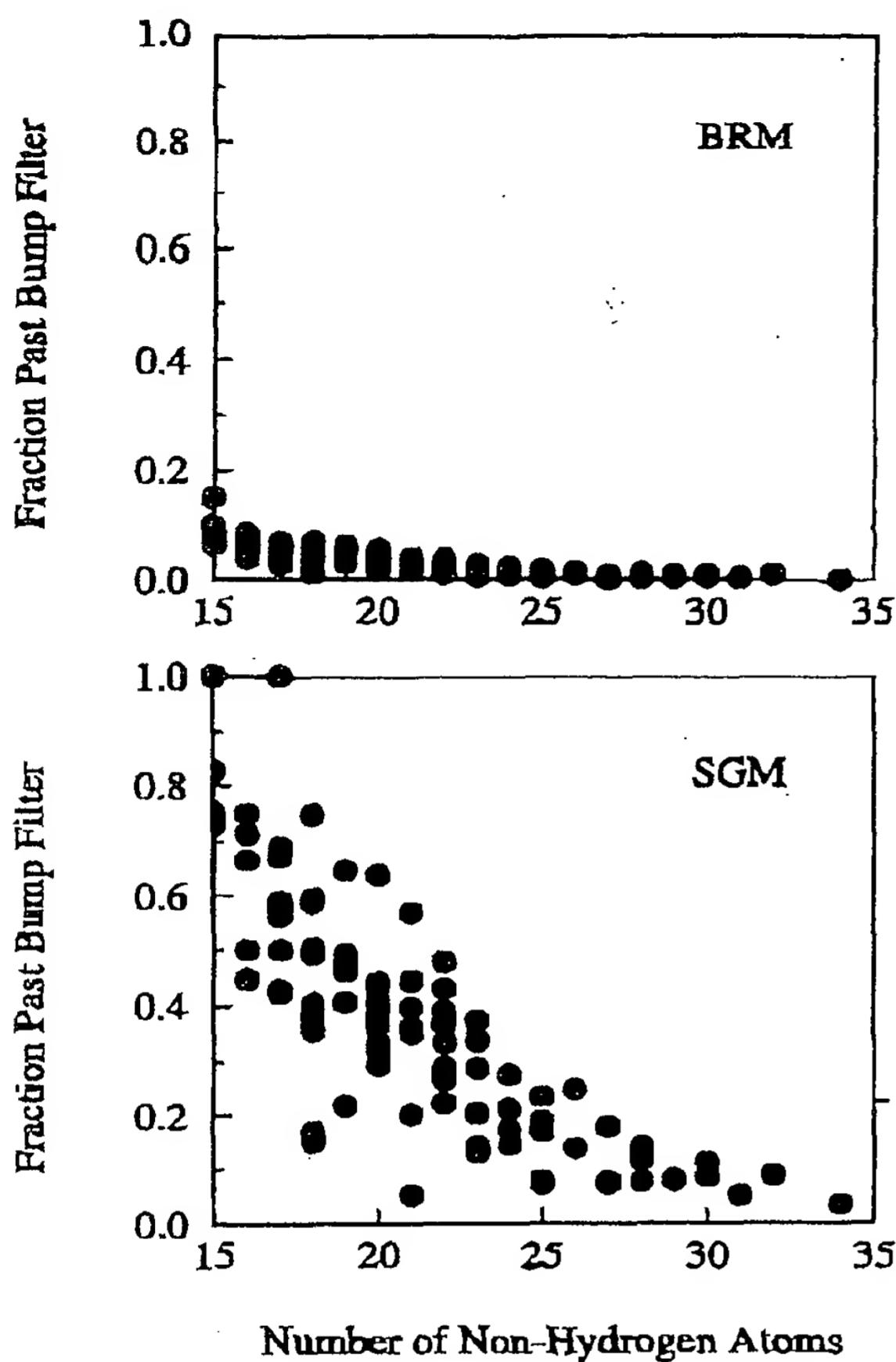
FIGURE 9. How molecule size affects bump filtering. In a given docking run, we compute the fraction of orientations that pass the bump filter for each molecule, and then plot this value against the size of the molecule. (Top) BRM method. Average filter rate is 0.031. (Bottom) SGM method. Average filter rate is 0.36.

in common. Therefore, distance matching will spend more time on molecules that are complementary in shape to the receptor. Thus, the BRM method benefits from using distance matching to determine the number of orientations to try simply because it will bias its efforts toward the larger and/or more complementary molecules.

Third, an orientation from random matching is, on average, not as good as one from distance matching. We can check the quality of these orientations by examining how they survive the bump filter as depicted in Figure 9. The orientations from distance matching are at least ten times more likely to make it past this filter, so they are indeed

superior. To understand why this superiority does not translate directly into faster docking, we must consider the computational bottleneck of the current implementation of DOCK. If the early docking steps—matching, orienting, and bump filtering—were to consume a dominant portion of the total cpu time, then SGM would be up to ten times faster than BRM. On the other hand, if the later docking steps—scoring and optimization—were to consume a dominant portion of the total cpu time, then BRM would be equivalent in speed to SGM. With the current implementation of the minimizer, the scoring and optimization steps indeed consume an overwhelming portion of cpu time (> 90%), so the ability of distance matching to form high-quality orientations is for now unrewarded. BRM is kept competitive because, in concert with heavy pruning by the bump filter, it forms orientations that are suitable starting points for optimization.

BRM VERSUS BGM

The fact that BRM outperforms BGM by up to a factor of two points out the critical importance of how processing time is allocated among the different molecules in the database. For maximum efficiency, a search algorithm must bias its efforts toward the most highly ranked molecules. BGM indeed has such a bias built in, but the heuristics have the effect of reducing the magnitude of bias inherent to the unrestrained matching found in SGM and BRM.

FUTURE DIRECTIONS

It appears sensible to incorporate the single graph matching technique into version 4.0 of DOCK, because its results are of high quality and its potential for speed improvement is high. The speed gains might arise from additional orientation filtering or by fundamental improvements in the optimization technique. The new matching algorithm will provide a solid algorithmic foundation on which to base further development of molecular docking, including the addition of sophisticated search constraints like chemical labels and critical clusters, as well as the explicit treatment of ligand flexibility.

Until a faster scoring and optimization method is implemented, it may be useful to preprocess the set of orientations generated by matching. For instance, an implementation of degeneracy checking has been tried in which similar orientations are

removed at the matching stage.[29] Because orienting is also relatively facile, it implies that degeneracy checking could instead be performed as an RMS deviation calculation between real orientations. Rarey et al. present a rapid RMSD evaluation technique based directly on the rotation matrices[17] that would be applicable. It might also be possible identify a unique property of orientations that lead to the best scores upon minimization (e.g., degree of surface overlap with receptor). We intend to further study the nature of orientations generated by BRM that pass the bump filter, to see why they manage to score so well upon minimization. Based upon this knowledge, a filter could be constructed that enriches the set of orientations generated by regular matching prior to minimization.

An additional avenue for improvement would be to try alternative methods to generate site points. The current shape-based site points may not be entirely consistent with a force-field-based scoring function.[21] Including force-field considerations during site point construction, might make the site points more relevant as points on which to position ligand atoms, thereby increasing the quality of docked orientations.

We are working to extend the current docking protocol to include ligand and receptor flexibility. Ligand flexibility can be incorporated in several ways. The most straightforward approach is to dock multiple conformations of the ligand separately.[21] A potentially more efficient method is to use distance geometry to build a ligand conformation that fits a subset of receptor site points.[10] Another viable option is the "divide and conquer" strategy, in which a flexible molecule is broken into rigid fragments, the fragments are docked independently, and the molecule is rebuilt from adjacent fragment orientations.[31-33] Limited receptor flexibility is being investigated during the step of score evaluation by superimposing multiple receptor conformers on a single score potential grid.[46]

Conclusion

We evaluated various search algorithms for automated molecular docking that range in complexity from purely random to site point-directed, graph-theoretical matching methods. Our basis of comparison was how quickly each docking method could correctly score and rank a database of molecules. Over a broad range of active site envi-

ronments, it is at least tenfold more efficient to use a collection of site points to describe the active site search volume than to use the smallest enclosing box. Using graph theoretical matching techniques boosts this relative efficiency higher. The bipartite graph matching used in the current DOCK version 3.5 improves efficiency up to 30-fold. The single graph matching proposed for DOCK version 4.0 improves efficiency up to 60-fold. Because single graph matching is not only faster, but also less complicated than the bipartite graph matching used in version 3.5, we feel it will be an important advance in docking technology.

References

1. A. Caflisch, P. Niederer, and M. Anliker, *Proteins*, 13, 223 (1992).

2. A. Di Nola, D. Roccatano, and H. J. C. Berendsen, *Proteins*, 19, 174 (1994).

3. D. S. Goodsell and A. J. Olson, *Proteins*, 8, 195 (1990).

4. J. B. Moon and W. J. Howe, *Proteins*, 11, 314 (1991).

5. R. Abagyan, M. Totrov, and D. Kuznetsov, *J. Comput. Chem.*, 15, 488 (1994).

6. A. Miranker and M. Karplus, *Proteins*, 11, 29 (1991).

7. A. R. Leach and I. D. Kuntz, *J. Comput. Chem.*, 13, 730 (1992).

8. A. R. Leach, *J. Mol. Biol.*, 235, 345 (1994).

9. M. Y. Mizutani, N. Tomioka, and A. Itai, *J. Mol. Biol.*, 243, 310 (1994).

10. A. S. Smellie, G. M. Crippen, and W. G. Richards, *J. Chem. Inf. Comput. Sci.*, 31, 386 (1991).

11. R. S. Judson, E. P. Jaeger, and A. M. Treasurywala, *Theochem J. Mol. Struct.*, 114, 191 (1994).

12. G. Jones, P. Willett, and R. C. Glen, *J. Mol. Biol.*, 245, 43 (1995).

13. C. M. Oshiro, I. D. Kuntz, and J. S. Dixon, *J. Comput.-Aided Mol. Design*, 9, 113 (1995).

14. Y. P. Pang and A. P. Kozikowski, *J. Comput.-Aided Mol. Design*, 8, 683 (1994).

15. E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser, *Proc. Natl. Acad. Sci. USA*, 89, 2195 (1992).

16. D. Fischer, S. L. Lin, H. L. Wolfson, and R. Nussinov, *J. Mol. Biol.*, 248, 459 (1995).

17. M. Rarey, S. Wefing, and T. Lengauer, *J. Comput.-Aided Mol. Design*, 10, 41 (1996).

18. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, *J. Mol. Biol.*, 161, 269 (1982).

19. F. S. Kuhl, G. M. Crippen, and D. K. Friesen, *J. Comput. Chem.*, 5, 24 (1984).

20. M. C. Lawrence and P. C. Davis, *Proteins*, 12, 31 (1992).

21. M. D. Miller, S. K. Kearsley, D. J. Underwood, and R. P. Sheridan, *J. Comput.-Aided Mol. Design*, 8, 153 (1994).

22. H.-J. Bohm, *J. Comput.-Aided Mol. Design*, 8, 623 (1994).

23. I. D. Kuntz, *Science*, 257, 1078 (1992).

24. D. R. Ferro and J. Hermans, *Acta Crystallogr.*, A33, 345 (1977).

25. Private communication from C. M. Oshiro and G. Golub.

26. B. K. Shoichet, D. L. Bodian, and I. D. Kuntz, *J. Comput. Chem.*, 13, 380 (1992).

27. E. C. Meng, B. K. Shoichet, and I. D. Kuntz, *J. Comput. Chem.*, 13, 505 (1992).

28. E. C. Meng, D. A. Gschwend, J. M. Blaney, and I. D. Kuntz, *Proteins*, 17, 266 (1993).

29. D. A. Gschwend and I. D. Kuntz, *J. Comput.-Aided Mol. Design*, 10, 123 (1996).

30. C. Bron and J. Kerbosch, *Commun. ACM*, 16, 575 (1973).

31. R. L. DesJarlais, R. P. Sheridan, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan, *J. Med. Chem.*, 29, 2149 (1986).

32. M. B. Eisen, D. C. Wiley, M. Karplus, and R. E. Hubbard, *Proteins*, 19, 199 (1994).

33. H.-J. Bohm, *J. Comput.-Aided Mol. Design*, 6, 61 (1992).

34. R. L. DesJarlais and J. S. Dixon, *J. Comput.-Aided Mol. Design*, 8, 231 (1994).

35. Distributed by Molecular Design Ltd., San Leandro, CA.

36. A. Rusinko et al., *J. Chem. Inf. Comput. Sci.*, 29, 251 (1989).

37. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, 112, 535 (1977).

38. E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, F. H. Allen, G. Bergerhoff, and R. Seivers, Eds., Data Commission of the International Union of Crystallography, Bonn, 1987, pp. 107–132.

39. L. Tabernero, N. Verdaguer, M. Coll, I. Fita, G. A. Van der Marel, J. H. Van Boom, A. Rich, and J. Aymami, *Biochemistry*, 32, 8403 (1993).

40. S. E. Ealick, Y. S. Babu, C. E. Bugg, M. D. Erion, W. C. Guida, J. A. Montgomery, and J. A. Secrist III, *Proc. Natl. Acad. Sci. USA*, 88, 11540 (1991).

41. J. T. Bolin, D. J. Filman, D. A. Matthews, R. C. Hamlin, and J. Kraut, *J. Biol. Chem.*, 257, 13650 (1982).

42. J. N. Herron, X. He, M. L. Mason, E. W. Voss Jr., and A. B. Edmundson, *Proteins*, 5, 271 (1989).

43. J. Erickson, D. J. Neidhart, J. Van Drie, D. J. Kempf, X. C. Wang, D. W. Norbeck, J. J. Plattner, J. W. Rittenhouse, M. Turon, N. Wideburg, W. E. Kohlbrenner, R. Simmer, R. Helfrich, D. A. Paul, and M. Knigge, *Science*, 249, 527 (1990).

44. S. L. Lin, R. Nussinov, D. Fischer, and H. J. Wolfsen, *Proteins*, 18, 94 (1994).

45. C. A. Reynolds, R. C. Wade, and P. J. Goodford, *J. Mol. Graph.*, 7, 103 (1989).

46. R. M. A. Knegtel, I. D. Kuntz, and C. M. Oshiro, *J. Mol. Biol.*, 266, 424 (1997).

Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function

GARRETT M. MORRIS,[1] DAVID S. GOODSELL,[1]
ROBERT S. HALLIDAY,[2] RUTH HUEY,[1] WILLIAM E. HART,[3]
RICHARD K. BELEW,[4] ARTHUR J. OLSON[1]

[1] Department of Molecular Biology, MB-5, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037-1000
[2] Hewlett-Packard, San Diego, California
[3] Applied Mathematics Department, Sandia National Laboratories, Albuqurque, NM
[4] Department of Computer Science & Engineering, University of California, San Diego, La Jolla, CA

ABSTRACT: A novel and robust automated docking method that predicts the bound conformations of flexible ligands to macromolecular targets has been developed and tested, in combination with a new scoring function that estimates the free energy change upon binding. Interestingly, this method applies a Lamarckian model of genetics, in which environmental adaptations of an individual's phenotype are reverse transcribed into its genotype and become heritable traits (sic). We consider three search methods, Monte Carlo simulated annealing, a traditional genetic algorithm, and the Lamarckian genetic algorithm, and compare their performance in dockings of seven protein–ligand test systems having known three-dimensional structure. We show that both the traditional and Lamarckian genetic algorithms can handle ligands with more degrees of freedom than the simulated annealing method used in earlier versions of AUTODOCK, and that the Lamarckian genetic algorithm is the most efficient, reliable, and successful of the three. The empirical free energy function was calibrated using a set of 30 structurally known protein–ligand complexes with experimentally determined binding constants. Linear regression analysis of the observed binding constants in terms of a wide variety of structure-derived molecular properties was performed. The final model had a residual standard error of 9.11 kJ mol^{-1} (2.177 kcal mol^{-1}) and was chosen as the new energy

Introduction

A fast atom-based computational docking tool is essential to most techniques for structure-based drug design.[1,2] Reported techniques for automated docking fall into two broad categories: matching methods and docking simulation methods.[3] Matching methods create a model of the active site, typically including sites of hydrogen bonding and sites that are sterically accessible, and then attempt to dock a given inhibitor structure into the model as a rigid body by matching its geometry to that of the active site. The most successful example of this approach is DOCK,[4,5] which is efficient enough to screen entire chemical databases rapidly for lead compounds. The second class of docking techniques model the docking of a ligand to a target in greater detail: the ligand begins randomly outside the protein, and explores translations, orientations, and conformations until an ideal site is found. These techniques are typically slower than the matching techniques, but they allow flexibility within the ligand to be modeled and can utilize more detailed molecular mechanics to calculate the energy of the ligand in the context of the putative active site. They allow computational chemists to investigate modifications of lead molecules suggested by the chemical intuition and expertise of organic synthetic chemists.

AUTODOCK[6,7] is an example of the latter, more physically detailed, flexible docking technique. Previous releases of AUTODOCK combine a rapid grid-based method for energy evaluation,[8,9] precalculating ligand–protein pairwise interaction energies so that they may be used as a look-up table during simulation, with a Monte Carlo simulated annealing search[10,11] for optimal conformations of ligands. AUTODOCK has been applied with great success in the prediction of bound conformations of enzyme–inhibitor complexes,[12,13] peptide–antibody complexes,[14] and even protein–protein interactions[15]; these and other applications have been reviewed elsewhere.[16]

We initiated the current work to remedy two limitations of AUTODOCK. (i) We have found that the simulated annealing search method performs well with ligands that have roughly eight rotatable bonds or less: problems with more degrees of freedom rapidly become intractable. This demanded a more efficient search method. (ii) AUTODOCK is often used to obtain unbiased dockings of flexible inhibitors in enzyme active sites: in computer-assisted drug-design, novel modifications of such lead molecules can be investigated computationally. Like many other computational approaches, AUTODOCK performs well in predicting relative quantities and rankings for series of similar molecules; however, it has not been possible to estimate in AUTODOCK whether a ligand will bind with a millimolar, micromolar, or nanomolar binding constant. Earlier versions of AUTODOCK used a set of traditional molecular mechanics force-field parameters that were not directly correlated with observed binding free energies; hence, we needed to develop a force field that could be used to predict such quantities.

Molecular docking is a difficult optimization problem, requiring efficient sampling across the entire range of positional, orientational, and conformational possibilities. Genetic algorithms (GA) fulfill the role of global search particularly well, and are increasingly being applied to problems that suffer from combinatorial explosions due to their many degrees of freedom. Both canonical genetic algorithms[17–21] and evolutionary programming methods[22] have been shown to be successful in both drug design and docking.

In this report, we describe two major advances that are included in the new release of AUTODOCK, version 3.0. The first is the addition of three new search methods: a genetic algorithm; a local search method; and a novel, adaptive global–local search method based on Lamarckian genetics, the Lamarckian genetic algorithm (LGA). The second advance is an empirical binding free energy force field that allows the prediction of binding free energies, and hence binding constants, for docked ligands.

Methods

GENETIC ALGORITHMS

Genetic algorithms[23] use ideas based on the language of natural genetics and biological evolution.[24] In the case of molecular docking, the particular arrangement of a ligand and a protein can be defined by a set of values describing the translation, orientation, and conformation of the ligand with respect to the protein: these are the ligand's state variables and, in the GA, each state variable corresponds to a gene. The ligand's state corresponds to the genotype, whereas its atomic coordinates correspond to the phenotype. In molecular docking, the fitness is the total interaction energy of the ligand with the protein, and is evaluated using the energy function. Random pairs of individuals are mated using a process of crossover, in which new individuals inherit genes from either parent. In addition, some offspring undergo random mutation, in which one gene changes by a random amount. Selection of the offspring of the current generation occurs based on the individual's fitness: thus, solutions better suited to their environment reproduce, whereas poorer suited ones die.

A variety of approaches have been adopted to improve the efficiency of the genetic algorithm. Classical genetic algorithms represent the genome as a fixed-length bit string, and employ binary crossover and binary mutation to generate new individuals in the population. Unfortunately, in many problems, such binary operators can generate values that are often outside the domain of interest, leading to gross inefficiencies in the search. The use of real encodings helps to limit the genetic algorithm to reasonable domains. Alternative genetic algorithms have been reported[25] that employ more complicated representations and more sophisticated operators besides crossover and mutation. Some of these retain the binary representation, but must employ decoders and repair algorithms to avoid building illegal individuals from the chromosome, and these are frequently computationally intensive. However, the search performance of the genetic algorithm can be improved by introducing a local search method.[26,27]

HYBRID SEARCH METHODS IN AUTODOCK

Earlier versions of AutoDock used optimized variants of simulated annealing.[6,7] Simulated annealing may be viewed as having both global and local search aspects, performing a more global search early in the run, when higher temperatures allow transitions over energy barriers separating energetic valleys, and later on performing a more local search when lower temperatures place more focus on local optimization in the current valley. AutoDock 3.0 retains the functionality of earlier versions, but adds the options of using a genetic algorithm (GA) for global searching, a local search (LS) method to perform energy minimization, or a combination of both, and builds on the work of Belew and Hart[27,28] The local search method is based on that of Solis and Wets,[29] which has the advantage that it does not require gradient information about the local energy landscape, thus facilitating torsional space search. In addition, the local search method is adaptive, in that it adjusts the step size depending upon the recent history of energies: a user-defined number of consecutive failures, or increases in energy, cause the step size to be doubled; conversely, a user-defined number of consecutive successes, or decreases in energy, cause the step size to be halved. The hybrid of the GA method with the adaptive LS method together form the so-called Lamarckian genetic algorithm (LGA), which has enhanced performance relative to simulated annealing and GA alone,[21,26] and is described in detail later. Thus, the addition of these new GA-based docking methods enhances AutoDock, and allows problems with more degrees of freedom to be tackled. Furthermore, it is now possible to use the same force field as is used in docking to perform energy minimization of ligands.

IMPLEMENTATION

In our implementation of the genetic algorithm, the chromosome is composed of a string of real-valued genes: three Cartesian coordinates for the ligand translation; four variables defining a quaternion specifying the ligand orientation; and one real-value for each ligand torsion, in that order. Quaternions are used to define the orientation[30] of the ligand, to avoid the gimbal lock problem experienced with Euler angles.[31] The order of the genes that encode the torsion angles is defined by the torsion tree created by AutoTors, a preparatory program used to select rotatable bonds in the ligand. Thus, there is a one-to-one mapping from the ligand's state variables to the genes of the individual's chromosome.

The genetic algorithm begins by creating a random population of individuals, where the user

defines the number of individuals in the population. For each random individual in the initial population, each of the three translation genes for x, y, and z is given a uniformly distributed random value between the minimum and maximum x, y, and z extents of the grid maps, respectively; the four genes defining the orientation are given a random quaternion, consisting of a random unit vector and a random rotation angle between $-180°$ and $+180°$; and the torsion angle genes, if any, are given random values between $-180°$ and $+180°$. Furthermore, a new random number generator has been introduced that is hardware-independent.[32] It is used in the LS, GA, and LGA search engines, and allows results to be reproduced on any hardware platform given the same seed values. The creation of the random initial population is followed by a loop over generations, repeating until the maximum number of generations or the maximum number of energy evaluations is reached, whichever comes first. A generation consists of five stages: mapping and fitness evaluation, selection, crossover, mutation, and elitist selection, in that order. In the Lamarckian GA, each generation is followed by local search, being performed on a user-defined proportion of the population. Each of these stages is discussed in more detail in what follows.

Mapping translates from each individual's genotype to its corresponding phenotype, and occurs over the entire population. This allows each individual's *fitness* to be evaluated. This is the sum of the intermolecular interaction energy between the ligand and the protein, and the intramolecular interaction energy of the ligand. The physicochemical nature of the energy evaluation function is described in detail later. Every time an individual's energy is calculated, either during global or local search, a count of the total number of energy evaluations is incremented.

This is followed, in our implementation, by proportional *selection* to decide which individuals will reproduce. Thus, individuals that have better-than-average fitness receive proportionally more offspring, in accordance with:

$$n_o = \frac{f_w - f_i}{f_w - \langle f \rangle} \qquad f_w \neq \langle f \rangle$$

where n_o is the integer number of offspring to be allocated to the individual; f_i is the fitness of the individual (i.e., the energy of the ligand); f_w is the fitness of the worst individual, or highest energy,

in the last N generations (i.e., N is a user-definable parameter, typically 10); and $\langle f \rangle$ is the mean fitness of the population. Because the worst fitness, f_w, will always be larger than either f_i or $\langle f \rangle$, except when $f_i = f_w$, then for individuals that have a fitness lower than the mean, $f_i < \langle f \rangle$, the numerator in this equation, $f_w - f_i$, will always be greater than the denominator $f_w - \langle f \rangle$, and thus such individuals will be allocated at least one offspring, and thus will be able to reproduce. AUTODOCK checks for $f_w = \langle f \rangle$ beforehand, and if true, the population is assumed to have converged, and the docking is terminated.

Crossover and *mutation* are performed on random members of the population according to user-defined rates of crossover and mutation. First, crossover is performed. Two-point crossover is used, with breaks occurring only between genes, never within a gene—this prevents erratic changes in the real values of the genes. Thus, both parents' chromosomes would be broken into three pieces at the same gene positions, each piece containing one or more genes; for instance, ABC and abc. The chromosomes of the resulting offspring after two-point crossover would be AbC and aBc. These offspring replace the parents in the population, keeping the population size constant. Crossover is followed by mutation; because the translational, orientational, and torsional genes are represented by real variables, the classical bit-flip mutation would be inappropriate. Instead, mutation is performed by adding a random real number that has a Cauchy distribution to the variable, the distribution being given by:

$$C(\alpha, \beta, x) = \frac{\beta}{\pi \left(\beta^2 + (x - \alpha)^2 \right)}$$

$$\alpha \geq 0, \; \beta > 0, \; -\infty < x < \infty$$

where α and β are parameters that affect the mean and spread of the distribution. The Cauchy distribution has a bias toward small deviates, but, unlike the Gaussian distribution, it has thick tails that enable it to generate large changes occasionally.[26]

An optional user-defined integer parameter *elitism* determines how many of the top individuals automatically survive into the next generation. If the elitism parameter is non-zero, the new population that has resulted from the proportional selection, crossover, and mutation is sorted according to its fitness; the fitness of new individuals

having resulted from crossover and/or mutation is calculated as necessary. Because populations are implemented as heaps, selection of the best n individuals is efficient.

The genetic algorithm iterates over generations until one of the termination criteria is met. At the end of each docking, AutoDock reports the fitness (the docked energy), the state variables, and the coordinates of the docked conformation, and also the estimated free energy of binding. AutoDock performs the user-specified number of GA dockings, and then carries out conformational cluster analysis on the docked conformations to determine which are similar, reporting the clusters ranked by increasing energy.

LAMARCKIAN GENETIC ALGORITHM

The vast majority of genetic algorithms mimic the major characteristics of Darwinian evolution and apply Mendelian genetics. This is illustrated on the right-hand side of Figure 1 (note the one-way transfer of information from the genotype to the phenotype). However, in those cases where an inverse mapping function exists (i.e., one which yields a genotype from a given phenotype), it is possible to finish a local search by replacing the individual with the result of the local search; see the left-hand side of Figure 1. This is called the Lamarckian genetic algorithm (LGA), and is an allusion to Jean Batiste de Lamarck's (discredited) assertion that phenotypic characteristics acquired during an individual's lifetime can become heritable traits.[33]

The most important issues arising in hybrids of local search (LS) techniques with the GA revolve around the *developmental mapping*, which transforms genotypic representations into phenotypic ones.[26] The genotypic space is defined in terms of the genetic operators—mutation and crossover in our experiments—by which parents of one generation are perturbed to form their children. The phenotypic space is defined directly by the problem, namely, the energy function being optimized. The local search operator is a useful extension of GA global optimization when there are local "smoothness" characteristics (continuity, correlation, etc.) of the fitness function that local search can exploit. In hybrid GA + LS optimizations, the result of the LS is always used to update the fitness associated with an individual in the GA selection algorithm. If, and only if, the developmental mapping function is *invertible*, will the Lamarckian option—
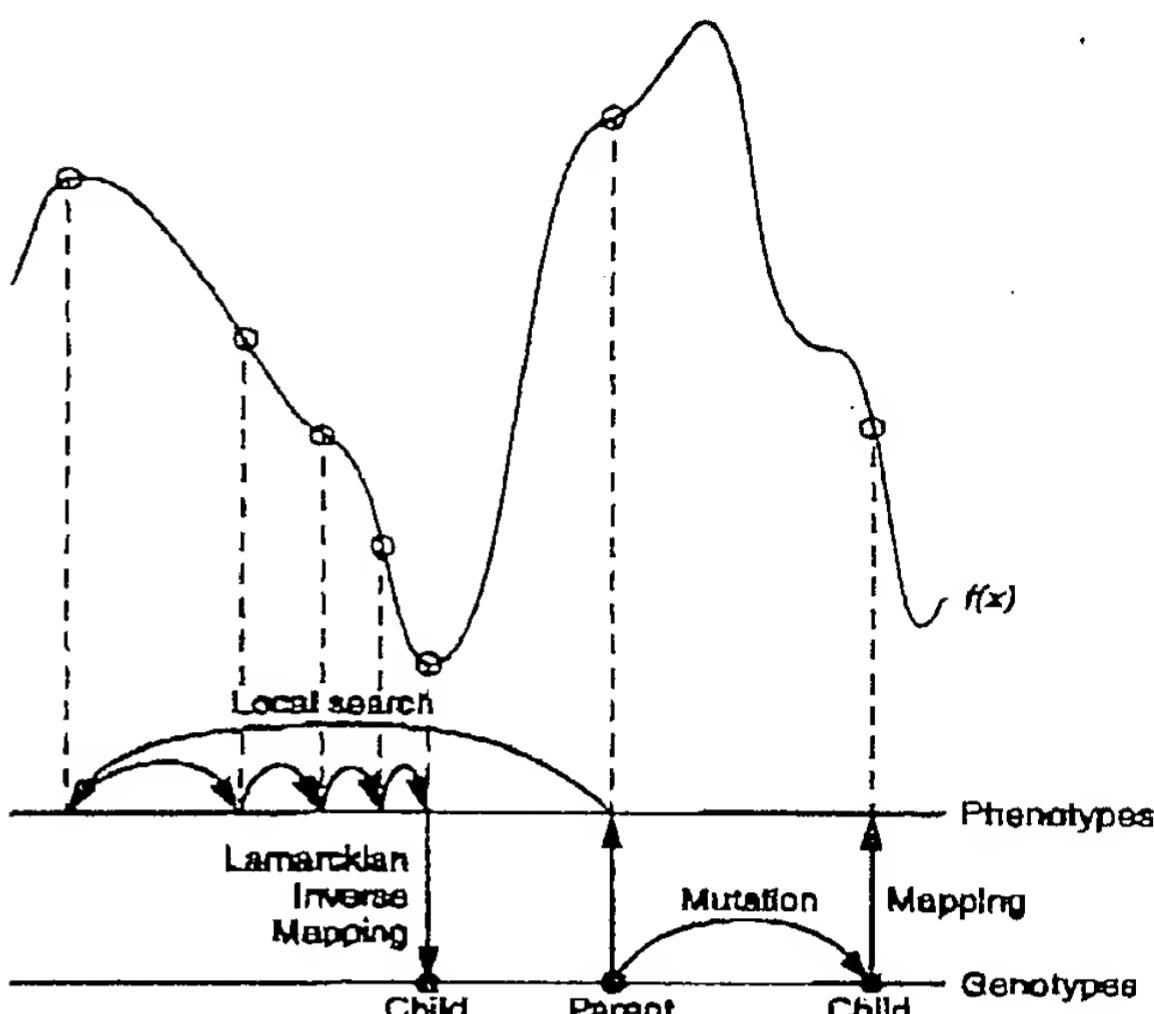
FIGURE 1. This figure illustrates genotypic and phenotypic search, and contrasts Darwinian and Lamarckian search.[27] The space of the genotypes is represented by the lower horizontal line, and the space of the phenotypes is represented by the upper horizontal line. Genotypes are mapped to phenotypes by a developmental mapping function. The fitness function is $f(x)$. The result of applying the genotypic mutation operator to the parent's genotype is shown on the right-hand side of the diagram, and has the corresponding phenotype shown. Local search is shown on the left-hand side. It is normally performed in phenotypic space and employs information about the fitness landscape. Sufficient iterations of the local search arrive at a local minimum, and an inverse mapping function is used to convert from its phenotype to its corresponding genotype. In the case of molecular docking, however, local search is performed by continuously converting from the genotype to the phenotype, so inverse mapping is not required. The genotype of the parent is replaced by the resulting genotype, however, in accordance with Lamarckian principles.

converting the phenotypic result of LS back into its corresponding genotype become possible.

In our case, the fitness or energy is calculated from the ligand's coordinates, which together form its phenotype. The genotypic representation of the ligand, and its mutation and crossover operators, have already been described. The developmental mapping simply transforms a molecule's genotypic state variables into the corresponding set of atomic coordinates. A novel feature of this application of hybrid global–local optimization is that the

Solis and Wets LS operator searches through the genotypic space rather than the more typical phenotypic space. This means that the developmental mapping does not need to be inverted. Nonetheless, this molecular variation of the genetic algorithm still qualifies as Lamarckian, because any "environmental adaptations" of the ligand acquired during the local search will be inherited by its offspring.

At each generation, it is possible to let a user-defined fraction of the population undergo such a local search. We have found improved efficiency of docking with local search frequencies of just 0.06, although a frequency of 1.00 is not significantly more efficient.[26] Both the canonical and a slightly modified version of the Solis and Wets method have been implemented. In canonical Solis and Wets, the same step size would be used for every gene, but we have improved the local search efficiency by allowing the step size to be different for each type of gene: a change of 1 Å (1 Å = 10^{-10} m) in a translation gene could be much more significant than a change of 1° in a rotational or torsional gene. In the docking experiments presented here, the translational step size was 0.2 Å, and the orientational and torsional step sizes were 5°.

In the Lamarckian genetic algorithm, genotypic mutation plays a somewhat different role than it does in traditional genetic algorithms. Traditionally, mutation plays the role of a local search operator, allowing small, refining moves that are not efficiently made by crossover and selection alone. With the explicit local search operator, however, this role becomes unnecessary, and is needed only for its role in replacing alleles that might have disappeared through selection. In LGA, mutation can take on a more exploratory role. The Cauchy deviates are a compromise between radical jumps to arbitrary sections of the conformation space and detailed exploration of the local topography.

DERIVATION OF THE EMPIRICAL BINDING FREE ENERGY FUNCTION

The study of molecular structure underpins much of computational molecular biology. There are several established methods for performing molecular mechanics and molecular dynamics, notably AMBER,[34,35] CHARMM,[36] DISCOVER,[37] ECEPP,[38] and GROMOS.[39] Many of these traditional force fields model the interaction energy of a molecular system with terms for dispersion/repulsion,[40] hydrogen bonding,[41] electrostatics,[42–45] and devia-

tion from ideal bond lengths and bond angles. These methods are excellent for studying molecular processes over time, for optimizing bound conformations, and for performing free energy perturbation calculations between molecules with a single atom change,[46] but they often require considerable investments of computer time and, unfortunately, these approaches tend to perform less well in ranking the binding free energies of compounds that differ by more than a few atoms. What is needed is an empirical relationship between molecular structure and binding free energy.

The first thoroughly established linear free energy relationship was observed by Hammett as early as 1933, and reported in 1937.[47] It was used to relate structure and reactivity of small organic molecules on a quantitative basis. Hammett was able to derive substituent constants and reaction constants that could then be used to calculate rate constants and equilibrium constants for a specific reaction of a specific compound. It could be said that Hammett's work was the forerunner of modern-day quantitative structure–activity relationships (QSAR), pioneered by Hansch and coworkers in the 1960s. Here it is assumed that the sum of the steric, electronic, and hydrophobic effects of substituents in a compound determines its biological activity; see, for example, Fujita,[48] Hansch,[49] and more recently Selassie et al.[50]

Current structure-based scoring functions seek to remedy some of the deficiencies of traditional force fields by developing empirical free energy functions that reproduce observed binding constants. Most of these approaches use an expanded "master equation" to model the free energy of binding, adding entropic terms to the molecular mechanics equations[51]:

$$\Delta G = \Delta G_{vdw} + \Delta G_{hbond} + \Delta G_{elec} + \Delta G_{conform} + \Delta G_{tor} + \Delta G_{sol}$$

where the first four terms are the typical molecular mechanics terms for dispersion/repulsion, hydrogen bonding, electrostatics, and deviations from covalent geometry, respectively; ΔG_{tor} models the restriction of internal rotors and global rotation and translation; and ΔG_{sol} models desolvation upon binding and the hydrophobic effect (solvent entropy changes at solute–solvent interfaces). This latter term is the most challenging. Most workers use variants of the method of Wesson and Eisen-

berg.[52] calculating a desolvation energy based on the surface area buried upon complex formation, with the area of each buried atom being weighted by an atomic solvation parameter. Böhm built on earlier work with the *de novo* inhibitor design program LUDI,[53] and used linear regression to calibrate a similar function against a set of 45 diverse protein–ligand complexes with published binding constants.[54] The final function predicted binding constants for a set of test complexes with a standard deviation equivalent to about a factor of 25 in binding constant: more than sufficient to rank inhibitors with millimolar, micromolar, and nanomolar binding constants. Jain devised a continuous, differentiable scoring function,[55] which is, in essence, very similar to that of Böhm, but based on non-physical pairwise potentials using Gaussians and sigmoidal terms.

We have implemented a similar approach using the thermodynamic cycle of Wesson and Eisenberg.[52] The function includes five terms:

$$\Delta G = \Delta G_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right)$$

$$+ \Delta G_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right)$$

$$+ \Delta G_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}}$$

$$+ \Delta G_{tor} N_{tor}$$

$$+ \Delta G_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{(-r_{ij}^2/2\sigma^2)} \quad (1)$$

where the five ΔG terms on the right-hand side are coefficients empirically determined using linear regression analysis from a set of protein–ligand complexes with known binding constants, shown in Table I. The summations are performed over all pairs of ligand atoms, *i*, and protein atoms, *j*, in addition to all pairs of atoms in the ligand that are separated by three or more bonds.

The *in vacuo* contributions include three interaction energy terms, used in previous versions of AUTODOCK: a Lennard–Jones 12-6 dispersion/repulsion term; a directional 12–10 hydrogen bonding term, where E(t) is a directional weight based on the angle, *t*, between the probe and the target atom[9]; and a screened Coulombic electrostatic potential.[56] Each of these terms, including their parameterization, have already been described.[7]

A measure of the unfavorable entropy of ligand binding due to the restriction of conformational

degrees of freedom is added to the *in vacuo* function. This term is proportional to the number of sp³ bonds in the ligand, N_{tor}.[54] We investigated variants that included and excluded methyl, hydroxyl, and amine rotors.

In the development of an empirical free energy function for AUTODOCK, the desolvation term was most challenging, because AUTODOCK uses a grid-based method for energy evaluation, and most published solvation methods are based on surface area calculations. We investigated two different methods of calculating the desolvation energy term. The first of these methods was based on estimating atom-by-atom contributions to the interfacial molecular surface area between the ligand and the protein using the difference in the surface areas of the complex and the unbound protein and unbound ligand. Both the solvent-accessible and solvent-excluded surface areas were considered, being calculated with MSMS,[57] a fast and reliable program that computes analytical molecular surfaces. Unfortunately, there can be significant errors in the value of the interfacial solvent-accessible surface areas, due to the "collar" of accessible surface that surrounds the ligand–protein interface in the complex. We also tested seven variants of the .pairwise, volume-based method of Stouten et al.[58]: this method has the advantage that it is consistent with the pre-calculated affinity grid formulation used by AUTODOCK. For each atom in the ligand, fragmental volumes of surrounding protein atoms are weighted by an exponential function and then summed, evaluating the percentage of volume around the ligand atom that is occupied by protein atoms. This percentage is then weighted by the atomic solvation parameter of the ligand atom to give the desolvation energy. The full method may be broken into four separate components: burial of apolar atoms in the ligand, burial of apolar protein atoms, burial of polar and charged atoms in the ligand, and burial of polar and charged protein atoms. Great success has also been reported in using simply the amount of hydrophobic surface area buried upon complexation as a measure of the "hydrophobic effect,"[54] so we tested several formulations that included only the volume lost around ligand carbon atoms. The burial of polar atoms caused particular problems, as discussed in what follows. Apart from the volume-based method, we tested a simpler formulation for the solvent transfer of polar atoms; that is, a constant term corresponding to the favorable free energy of interaction of a polar atom with solvent

TABLE I.
Protein–Ligand Complexes Used to Calibrate Empirical Free Energy Function, Along with Brookhaven Protein Data Bank (PDB) Accession Codes and Binding.

Protein–ligand complex	PDB code	Log(K_i)[a]
Concanavalin A / α-methyl-D-mannopyranoside	4cna	2.00
Carboxypeptidase A / glycyl-L-tyrosine	3cpa	3.88
Carboxypeptidase A / phosphonate ZAA=P=(O)F	6cpa	11.52
Cytochrome P-450$_{cam}$ / camphor	2cpp	6.07
Dihydrofolate reductase / methotrexate	4dfr	9.70
α-Thrombin / benzamidine	1dwb	2.92
Endothiapepsin / H-256	zer6	7.22
ε-Thrombin / MQPA	1etr	7.40
ε-Thrombin / NAPAP	1ets	8.52
ε-Thrombin / 4-TAPAP	1ett	6.19
FK506-binding protein (FKBP) / immunosuppressant FK506	1fkf	9.70
D-Galactose / D-glucose binding protein / galactose	2gbp	7.60
Hemagglutinin / sialic acid	4hmg	2.55
HIV-1 Protease / A78791	1hvj	10.46
HIV-1 Protease / MVT101	4hvp	6.15
HIV-1 Protease / acylpepstatine	5hvp	5.96
HIV-1 Protease / XK263	1hvr	9.51
Fatty-acid-binding protein / $C_{15}COOH$	2ifb	5.43
Myoglobin (ferric) / imidazole	1mbi	1.88
McPC603 / phosphocholine	2mcp	5.23
β-Trypsin / benzamidine	3ptb	4.74
Retinol-binding protein / retinol	1rbp	6.72
Thermolysin / Leu-hydroxylamine	4tln	3.72
Thermolysin / phosphoramidon	1tlp	7.55
Thermolysin / n-(1-carboxy-3-phenylpropyl)-Leu-Trp	1tmn	7.30
Thermolysin / Cbz-Phe-p-Leu-Ala (ZFpLA)	4tmn	10.19
Thermolysin / Cbz-Gly-p-Leu-Leu (ZGpLL)	5tmn	8.04
Purine nucleoside phosphorylase (PNP) / guanine	1ulb	5.30
Xylose isomerase / CB3717	2xis	5.82
Triose phosphate isomerase (TIM) / 2-phosphoglycolic acid (PGA)	2ypi	4.82

[a] Adapted from Böhm.[54]

is estimated, and this is subtracted from the binding free energy.

Trilinear interpolation is used to evaluate rapidly the intermolecular dispersion/repulsion energy, the hydrogen bonding energy, the electrostatic potential, and the solvation energy of each atom in the ligand, using grid maps that have been pre-calculated over the protein for each atom type in the ligand. In AUTODOCK 3.0, we have implemented a faster method of trilinear interpolation[59] than was available in earlier versions of AUTODOCK. Both methods are mathematically equivalent. The original implementation used 24 multiplications to perform each three-dimensional trilinear interpolation, but, by cascading seven one-dimensional interpolations, the number of multiplications has been reduced to 7.

Thirty protein–ligand complexes with published binding constants were used in the calibration of AUTODOCK's free energy function (Table I), and were chosen from the set of 45 used by Böhm,[54] omitting all complexes that he modeled (i.e., using only complexes for which crystallographic structures were available). One of the limitations of these binding constant data is that the conditions under which they were determined vary, which intrinsically limits the accuracy of our best model. We converted between the inhibition constant, K_i, and the observed free energy change of binding, ΔG_{obs}, using the equation:

$$\Delta G_{obs} = RT \ln K_i$$

where R is the gas constant, 1.987 cal K^{-1} mol^{-1}, and T is the absolute temperature, assumed to be

room temperature, 298.15 K.[60] Note that this equation lacks a minus sign because the inhibition constant is defined for the dissociation reaction, $EI \rightleftharpoons E + I$,[61] whereas ΔG_{obs} refers to the opposite process of binding, $E + I \rightleftharpoons EI$; where E is the enzyme and I is the inhibitor.

To remove any steric clashes in the crystallographic complexes, each ligand was optimized using AutoDock's new Solis and Wets local minimization technique described earlier, but with the previously reported force field.[7] The separate contributions from the hydrogen bonding, dispersion/repulsion, electrostatic, and solvation energies were evaluated. Empirical free energy coefficients for each of these terms were derived using linear regression in the S-PLUS software package,[62] and cross-validation studies were performed. In total, 900 different binding free energy models were tested: each linear model consisted of a van der Waals term, a hydrogen bonding term (one of 6 variants), an electrostatic term, a torsional entropy term (one of 5 variants), and a desolvation term (one of 15 variants). We also investigated whether the inclusion of a constant term improved the model. Six of the seven test systems used to test the docking procedure, which were originally used to test AutoDock, version 2.4,[7] were also in the training set of 30 protein–ligand complexes; therefore, to validate the chosen coefficients, linear regression was repeated for the set of 24 protein–ligand complexes, excluding the 6 overlapping test systems.

TESTING DOCKING METHODS

Seven protein–ligand complexes, with a range of complexity and chemical properties, were chosen from the Brookhaven Protein Data Bank[63,64] to compare the performance of the docking techniques (see Fig. 2). To facilitate comparison with the previous force field, we chose the same set of six test systems investigated earlier,[7] but added a harder docking problem to challenge all the search methods (see Table II). The simplest test cases were the β-trypsin/benzamidine and cytochrome P-450$_{cam}$/camphor complexes, which had small, rigid ligands. Interactions in the former are dominated by electrostatic interactions and hydrogen bonds to the substrate amidine, whereas the latter is dominated by hydrophobic interactions. McPC-603/phosphocholine and streptavidin/biotin were moderately flexible, and represented test systems having an intermediate level of difficulty. HIV-1

protease/XK263, hemagglutinin/sialic acid, and dihydrofolate reductase/methotrexate provided more difficult tests, with many rotatable bonds and diverse chemical characteristics.

We compared the performance of Monte Carlo simulated annealing (SA), the genetic algorithm (GA), and the Lamarckian genetic algorithm (LGA). The new empirical free energy function presented here was used for energy evaluation in all cases. Dockings were performed using approximately the same number of energy evaluations (~ 1.5 million), so each method could be judged given similar computational investments. The CPU time taken for a single docking varied from 4.5 to 41.3 minutes, on a 200-MHz Silicon Graphics MIPS 4400 with 128 MB of RAM, depending on the number of rotatable bonds and the number of atoms in the ligand.

At the end of a set of dockings, the docked conformations were exhaustively compared to one another to determine similarities, and were clustered accordingly. The user-defined root-mean-square positional deviation (rmsd) tolerance was used to determine if two docked conformations were similar enough to be included in the same cluster, and symmetrically related atoms in the ligand were considered. These clusters were ranked in order of increasing energy, by the lowest energy in each cluster. Ordinarily, the structure of the protein–ligand complex would not be known, so the criteria by which the dockings would be evaluated are the energies of the docked structures, and, in cases where there are several plausible, low-energy structures, the number of conformations in a conformationally similar cluster. Because one of our goals was to test the ability of the methods to reproduce known structures, we also compared the rmsd between the lowest energy docked structure and the crystallographic structure.

DOCKING- AND SEARCH-METHOD-SPECIFIC PARAMETERS

The proteins and ligands in the seven docking tests were treated using the united-atom approximation, and prepared using the molecular modeling program, Sybyl.[65] Only polar hydrogens were added to the protein, and Kollman united-atom partial charges were assigned. Unless stated otherwise, all waters were removed. Atomic solvation parameters and fragmental volumes were assigned to the protein atoms using a new AutoDock utility, AddSol. The grid maps were calculated using AutoGrid, version 3.0. In all seven protein–ligand
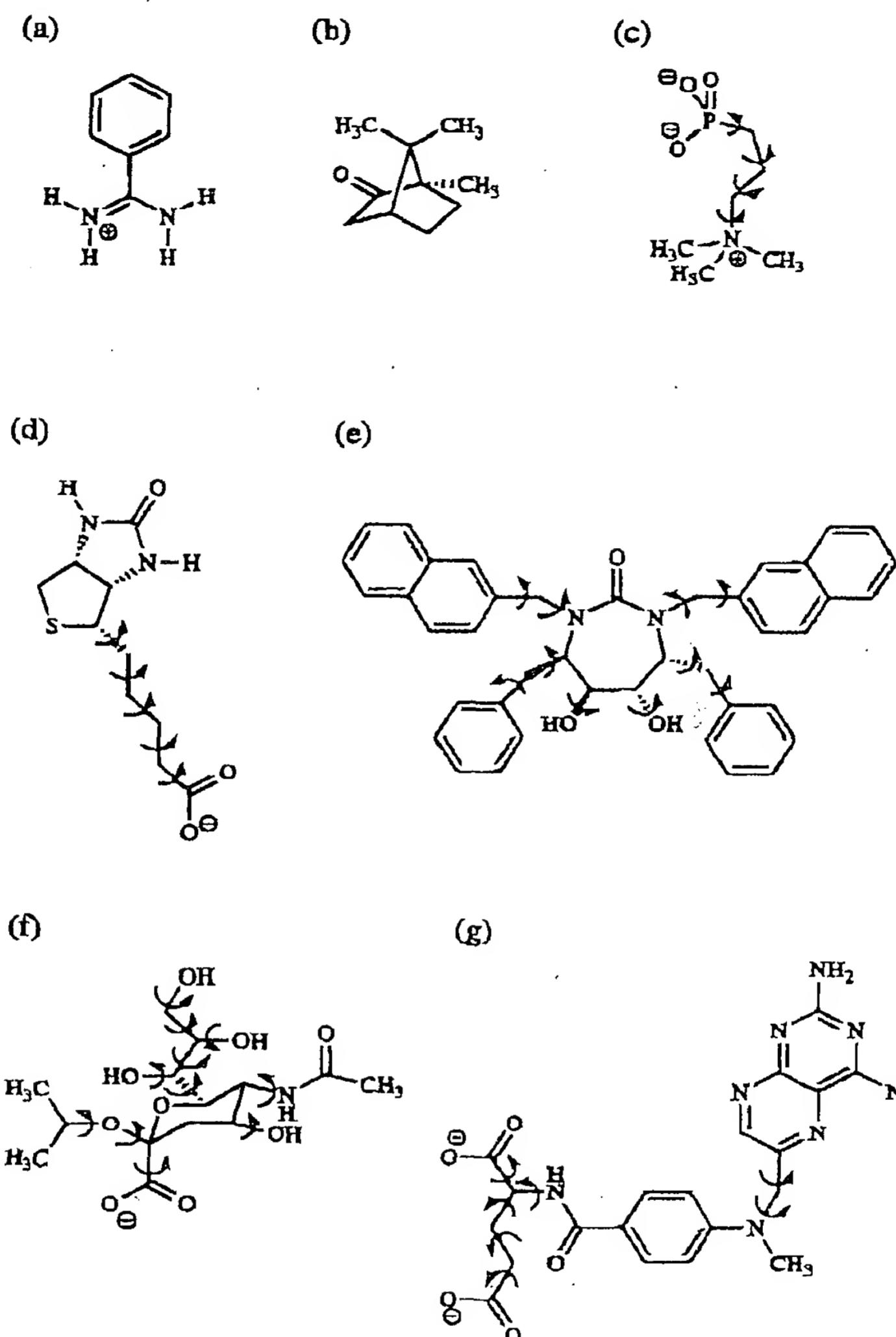
FIGURE 2. The seven ligands chosen for docking, showing the rotatable bonds as curly arrows: (a) benzamidine; (b) camphor; (c) phosphocholine; (d) biotin; (e) HIV-1 protease inhibitor XK-263; (f) isopropylated sialic acid; and (g) methotrexate. Note that two ligands, (e) and (f), contain hydroxyl rotors, which are not counted in the total number of torsional degrees of freedom; note also that cyclic rotatable bonds are excluded.

cases, we used grid maps with 61 × 61 × 61 points, a grid-point spacing of 0.375 Å, and, because the location of the ligand in the complex was known, the maps were centered on the ligand's binding site. The ligands were treated in SYBYL initially as all atom entities, that is, all hydrogens were added, then partial atomic charges were calculated using the Gasteiger–Marsili method.[66,67] AUTOTORS, an AUTODOCK utility, was used to define the rotatable bonds in the ligand, if any, and also to unite the

nonpolar hydrogens added by SYBYL for the partial atomic charge calculation. The partial charges on the nonpolar hydrogens were added to that of the hydrogen-bearing carbon also in AUTOTORS.

In all three search methods, 10 dockings were performed; in the analysis of the docked conformations, the clustering tolerance for the root-mean-square positional deviation was 0.5 Å, and the crystallographic coordinates of the ligand were used as the reference structure. For all three search

TABLE II.

X-Ray Crystal Structure Coordinates Used In Docking Experiments, Their Brookhaven Protein Data Bank Accession Codes and Resolution, Number of Rotatable Bonds In the Ligand, Number of Torsional Degrees of Freedom, Total Number of Degrees of Freedom, and Energy of Crystal Structure Using the Empirical Force Field Presented Here.

Protein–ligand complex	PDB code	Resolution (Å)	Reference	Number of rotatable bonds	N_{tor}[a]	Total number of degrees of freedom	Energy of crystal structure (kcal mol^{-1})
β-Trypsin / benzamidine	3ptb	1.7	69	0	0	7	−7.86
Cytochrome P-450$_{cam}$ / camphor	2cpp	1.63	70	0	0	7	−4.71
McPC-603 / Phosphocholine	2mcp	3.1	71	4	4	11	+5.48[b]
Streptavidin / biotin	1stp	2.6	73	5	5	12	−8.86
HIV-1 protease / XK263	1hvr	1.8	75	10	8	17	−18.62
Influenza hemagglutinin / sialic acid	4hmg	3.0	76	11	7	18	−4.71
Dihydrofolate reductase / methotrexate	4dfr	1.7	79	7	7	14	−13.64

[a] N_{tor}, is the number of torsional degrees of freedom used in the calculation of the predicted free energy change of binding, ΔG_{pred}. Note that this excludes rotatable bonds that only move hydrogens, such as hydroxyl, amino, and methyl groups.
[b] This energy is dominated by a large, positive contribution from C2 and O1 to the internal nonbonded energy, of +6.13 kcal mol^{-1}; these atoms are 2.26 Å apart.

methods, the step sizes were 0.2 Å for translations and 5° for orientations and torsions. These step sizes determined the amount by which a state variable could change when a move is made in simulated annealing and the relative size of mutation in the local search, whereas the α and β parameters determined the size of the mutation in the genetic algorithms, GA and LGA. The Cauchy distribution parameters were $\alpha = 0$ and $\beta = 1$. Note that in simulated annealing, random changes were generated by a uniformly distributed random number generator; in the Solis and Wets local search, by a normal distribution; and, in the genetic algorithm, by a Cauchy distribution. In the simulated annealing tests, the initial state of the ligand was chosen randomly by AuToDock. We used the optimal set of simulated annealing parameters that were determined from the schedule experiments described earlier.[7] These included an initial annealing temperature of 616 cal mol^{-1}, a linear temperature reduction schedule, 10 runs, 50 cycles, and a cycle-termination criterion of a maximum of 25,000 accepted steps or 25,000 rejected steps, whichever came first. The minimum energy state was used to begin the next cycle; the only exception was for 1hvr, where the initial annealing temperature was increased to 61,600 cal mol^{-1}. The maximum initial energy allowed was 0.0 kcal mol^{-1}, and the maximum number of retries was 1000, used to generate a low energy random initial state to begin each simulated annealing docking.

In the GA and LGA dockings, we used an initial population of random individuals with a population size of 50 individuals; a maximum number of 1.5×10^6 energy evaluations; a maximum number of generations of 27,000; an elitism value of 1, which was the number of top individuals that automatically survived into the next generation; a mutation rate of 0.02, which was the probability that a gene would undergo a random change; and a crossover rate of 0.80, which was the probability that two individuals would undergo crossover. Proportional selection was used, where the average of the worst energy was calculated over a window of the previous 10 generations. In the LGA dockings, the pseudo-Solis and Wets local search method was used, having a maximum of 300 iterations per local search; the probability of performing local search on an individual in the population was 0.06; the maximum number of consecutive successes or failures before doubling or halving the local search step size, ρ, was 4, in both cases; and the lower bound on ρ, the termination criterion for the local search, was 0.01.

Results and Discussion

CALIBRATION OF EMPIRICAL FREE ENERGY FUNCTION

Several linear regression models were tested for their ability to reproduce the observed binding

TABLE III.
Calibration of Empirical Free Energy Function.

Model[a]	Residual standard error	Multiple R^2	ΔG_{vdw}[b]	ΔG_{estat}	ΔG_{hbond}	ΔG_{tor}	ΔG_{solv}
A	2.324	0.9498	0.1795 (0.0263)	0.1133 (0.0324)	0.0166 (0.0625)	0.3100 (0.0873)	0.0101 (0.0585)
B	2.232	0.9537	0.1518 (0.0269)	0.1186 (0.0246)	0.0126 (0.0382)	0.3548 (0.0890)	0.1539 (0.1050)
C	2.177	0.9559	0.1485 (0.0237)	0.1146 (0.0238)	0.0656 (0.0558)	0.3113 (0.0910)	0.1711 (0.1035)

[a] Models differ in the formulation of the solvation term and the hydrogen bonding term. Model A: full volume-based solvation term and standard 10–12 hydrogen bonding, as in Eq. (1). Model B: apolar ligand atoms only in the solvation term, and standard 10–12 hydrogen bonding. Model C: apolar ligand atoms only in the solvation term, and the standard 10–12 hydrogen less the estimated average, as in Eq. (2).
[b] Values for the model coefficients, with standard deviations in parentheses.

constants of structurally characterized complexes. Table III shows the results for the three major candidates, and Figure 3 shows the correlation between the observed and the predicted binding free energies for the 30 protein–ligand complexes in the calibration set, using the chosen model (model C). Model A adds the full volume-based solvation method and the torsional restriction term to the original molecular mechanics force field. Model B simplifies the solvation method by evalu-
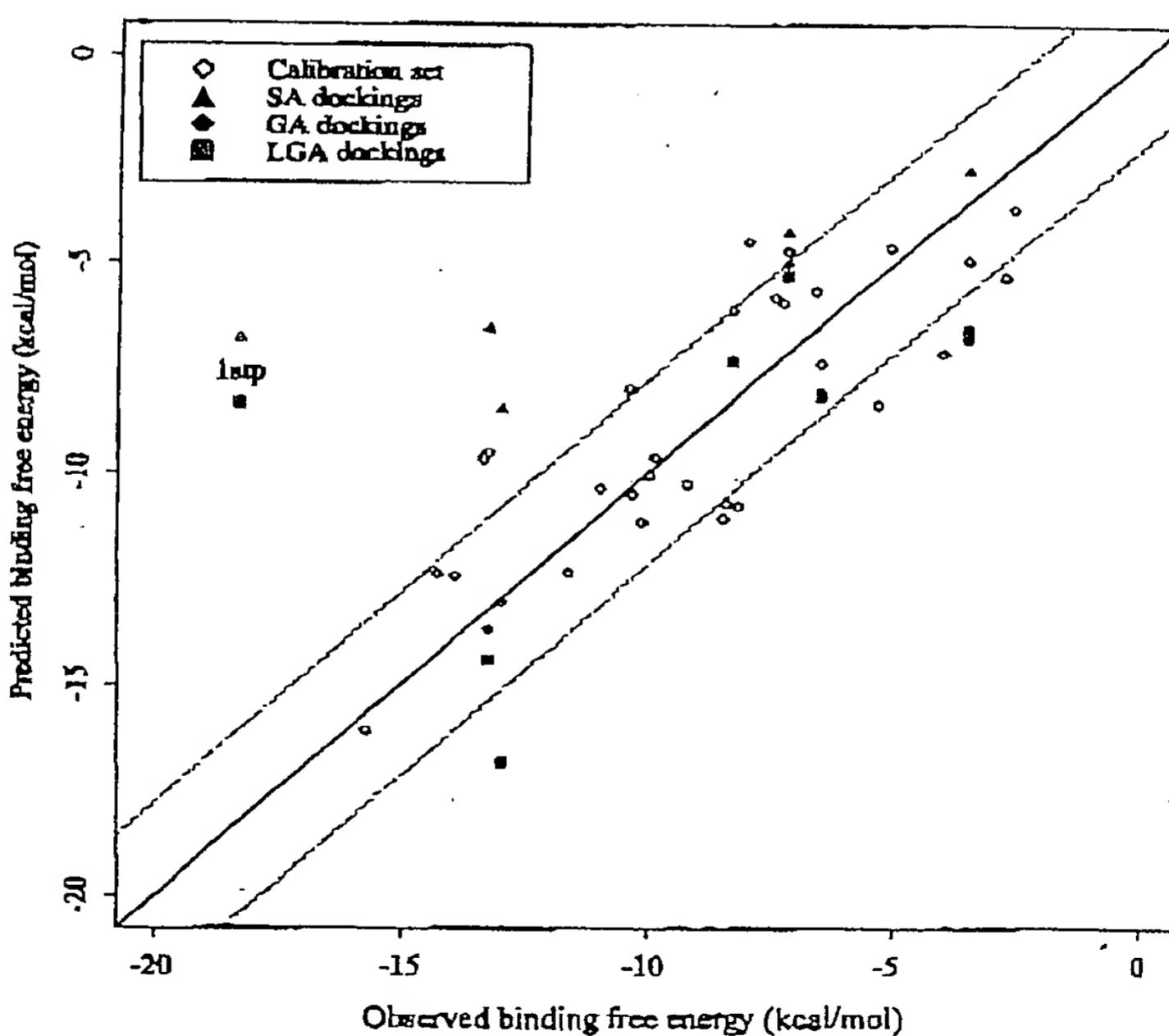
FIGURE 3. Predicted versus observed binding free energies for the calibration set and the docking tests. The solid line shows a perfect fit, and the dotted lines show one standard deviation above and below this. Hollow diamonds show the 30 protein–ligand complexes used in fitting the terms of the binding free energy function. Solid triangles show the results of the simulated annealing (SA) dockings, solid diamonds show the genetic algorithm (GA) dockings, and the solid squares show the Lamarckian genetic algorithm (LGA) dockings. Note the outlying biotin–streptavidin complex (1stp), where it is believed there are significant contributions to the binding free energy due to protein rearrangements.

ating the volume buried for only the carbon atoms in the ligand. Model C also uses only ligand carbon atoms in the desolvation calculation, and also adds a constant term to the hydrogen bonding function, modeling desolvation of polar atoms. Model C was chosen for incorporation into AUTODOCK 3.0, based on its better overall statistics, and on criteria discussed in what follows. The form of this free energy function is:

$$\Delta G = \Delta G_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right)$$
$$+ \Delta G_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} + E_{hbond} \right)$$
$$+ \Delta G_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}}$$
$$+ \Delta G_{tor} N_{tor}$$
$$+ \Delta G_{sol} \sum_{i_c,j} S_i V_j e^{(-r_{ij}^2/2\sigma^2)} \qquad (2)$$

where E_{hbond} is the estimated average energy of hydrogen bonding of water with a polar atom, and the summation in the solvation term is performed over all pairs consisting of only carbon atoms in the ligand, i, and atoms of all types, j, in the protein. Note that the internal or intramolecular interaction energy of the ligand is *not* included in the calculation of binding free energy; during docking, however, internal energy *is* included in the total docked energy, because changes in ligand conformation can affect the outcome of the docking, so this must be taken into consideration. We looked at linear regression models that did include the internal energy, and found that adding this term did not improve the model. The assumption made is that the internal energy of the ligand in solution and in the complex are the same. The energies used and reported by AUTODOCK should be distinguished: there are *docked energies*, which include the intermolecular and intramolecular interaction energies, and which are used during dockings; and *predicted free energies*, which include the intermolecular energy and the torsional free energy, and are only reported at the end of a docking. Because the intermolecular energy grid maps include the desolvation term, dockings using the new, empirical force field in AUTODOCK version 3.0 may be qualitatively different from results found using earlier versions.

Three coefficients, for dispersion/repulsion, electrostatics, and loss of torsional freedom were

very stable in the linear regression analysis, with consistent coefficient values in different formulations and reasonable standard deviations. In our best model, dispersion/repulsion energies, with parameters taken from AMBER,[54] were weighted by a factor of 0.1485, yielding an energy of about -0.2 kcal mol^{-1} for the most favorable atom–atom contacts. Electrostatics, modeled with a screened Coulomb potential,[55] were weighted by a factor of 0.1146, yielding an energy of about -1.0 kcal mol^{-1} for an ideal salt bridge. In the torsional restriction term, each torsional degree of freedom requires 0.3113 kcal mol^{-1}.

The major differences between models occurred with the interaction of the hydrogen bonding term and the desolvation term. Hydrogen bonding is modeled with a directional 12–10 potential.[68] We encountered a major problem when calibrating this hydrogen bonding function. Because the test set included only natural enzyme–ligand complexes, optimized by millions of years of evolution, hydrogen bonding groups in the ligands are nearly always paired with the appropriate hydrogen bonding group in the protein. Thus, the number of hydrogen bonds that the ligand forms in the complex and the number it forms with solvent when free in solution are approximately the same; that is, there is little change in the free energy of hydrogen bonding, and ΔG_{hbond} was evaluated to be approximately zero. Unfortunately, this provides no information on the cost of burying a hydrogen bonding group without forming a bond with the protein, and our data set did not include cases to evaluate this. Of course, the volume-based solvation method should account for this—the unfavorable polar contribution to the solvation energy should compensate for the favorable 12–10 hydrogen bonding energy. The linear regression, however, consistently returned coefficients that set the hydrogen bonding energy and desolvation energy to nearly zero, and increased the dispersion/repulsion term to compensate (see Model A in Table III). We chose an alternative formulation to resolve this problem.

We obtained the best results by separating the desolvation of polar atoms from the volume-based calculation. We assumed that the extent of hydrogen bonding in the complexes was roughly the same as the extent of hydrogen bonding in solution. The calculated hydrogen bonding energy, using the directional 12–10 hydrogen bonding function, was divided by the maximal number of possible hydrogen bonds, counting two for each oxygen atom and one for each polar hydrogen. For

the 30 complexes used in calibration, it was found that 36% of the maximum possible hydrogen bonding sites were actually utilized. Values of E_{hbond} ranging from 36% to 100% of the maximal well depth of 5 kcal mol^{-1} had little effect on the success of the formulation (data not shown), and a value of 36% was chosen. Optimized weights yielded an ideal hydrogen bonding energy in the complex of -0.328 kcal mol^{-1}, and the estimated average energy of each hydrogen bond in solution of -0.118. Because hydrogen bonding was modeled by this difference, the typical hydrogen bonding free energy of a complex was approximately zero, but there was a penalty of about 0.2 kcal mol^{-1} for oxygen and nitrogen atoms that did not form hydrogen bonds, driving the simulation toward docked conformations with maximal hydrogen bonding. We are currently exploring an appropriate data set for evaluating this formulation more rigorously.

The desolvation for carbon atoms in the ligand was evaluated using two different classes of atom type, aliphatic and aromatic, as in the original study.[58] The desolvation term was weighted by a factor of 0.1711 in our final empirical free energy force field, so a typical aliphatic carbon atom yields an energy of about -0.2 kcal mol^{-1} upon binding.

We cross-validated the free energy model in two ways. First, we investigated the influence of each member of the training set on the final coefficients of the model, by removing each one from the training set and calculating the coefficients from the remaining 29 complexes. We found that none of them had a strong effect on the final values of the coefficients.

We also performed a second kind of cross-validation of the free energy model, by performing Solis and Wets local search using AUTODOCK and the new free energy function, starting from the x-ray crystallographic conformations of each inhibitor in 20 HIV-1 protease-inhibitor complexes, to compare the resulting optimized conformations' predicted free energy change of binding, $\Delta G_{binding}$, with the experimentally determined values. These protease inhibitors were quite different, having from 7 to 28 torsions, and widely different side chains—charged, polar, and hydrophobic, and constituted a diverse test set. As can be seen from the results in Table IX, the correlation was very good, with an overall rmsd between the experimental and calculated values of $\Delta G_{binding}$ of 1.92 kcal mol^{-1}.

The final form of the free energy function may seem overparameterized, with additional weighting parameters added to a previously optimized parameterization. We retained the molecular mechanics formulation, however, specifically for its ability to model the distance dependence of each energetic term. This distance dependence (and angular dependence in hydrogen bonding) is essential for finding valid docked conformations, but the amount and resolution of the available protein–ligand data do not support a full re-pameterization of the functions.

DOCKING EXPERIMENTS

Because we are comparing different search methods, it is important to ensure that the methods are treated equally. It is therefore important that each search method be allowed approximately the same number of energy evaluations in a docking. The number of energy evaluations in a docking depends on the termination criteria, and because it is not possible to predict how many accepted or rejected steps the stochastic SA method will make at a given temperature, the number of evaluations varies in SA. The range was from 1.19×10^6 to 2.33×10^6, depending on the protein–ligand test system, even though the same parameters were used for the number of cycles, accepted steps and rejected steps. In the case of the GA dockings, the population was 50 and the number of generations was 27,000, which gave a total of 1.35×10^6 energy evaluations in a docking; thus, the GA dockings were terminated by reaching the maximum number of generations. In the case of the LGA dockings, 6% of the population underwent Lamarckian local search, each search consisting of 300 iterations and each iteration using an extra energy evaluation. Thus, the LGA dockings, even with the same population size and number of generations as the GA, were terminated by reaching the maximum number of energy evaluations, 1.50×10^6.

The results of the simulated annealing, genetic algorithm, and the Lamarckian genetic algorithm docking experiments are summarized in Tables IV, V and VI, respectively. The lowest energy docked structure found by each method is compared with the crystal structure of the ligand in Figure 4. The predicted change in free energy upon binding, ΔG_{pred}, for the lowest energy found by LGA is shown in Table VII, along with the experimentally observed change in free energy upon binding, ΔG_{obs}. In addition, the breakdown of the energy of the lowest energy docked conformation is shown, in terms of the intermolecular interaction energy,

TABLE IV.
Results of Simulated Annealing Dockings.[a]

PDB code	Number of clusters	Number in rank 1	Energy (kcal mol^{-1}) and rmsd (Å)						Number of energy evaluations
			Lowest energy	rmsd of lowest energy	Mean energy		Mean rmsd		
3ptb	5	6	−8.03	0.21	−7.84	(0.08)	0.50	(0.17)	2.01×10^6
2cpp	4	6	−7.29	0.81	−7.22	(0.03)	0.91	(0.30)	2.33×10^6
2mcp	10	1	−4.09	0.88	70.89	(2.10×10^2)	5.40	(4.80)	1.85×10^6
1stp	10	1	−8.48	1.27	−7.71	(0.66)	1.24	(0.95)	2.00×10^6
1hvr	10	1	−11.77	1.15	1.12×10^6	(3.36×10^5)	6.13	(2.61)	1.19×10^6
4hmg	10	1	−2.59	3.77	6.99×10^4	(1.52×10^5)	6.20	(2.94)	1.55×10^6
4dfr	10	1	−8.73	4.83	6.13×10^2	(1.96×10^3)	5.04	(1.74)	1.30×10^6

[a] The parameters used were 10 runs, 50 cycles, and a cycle-termination criterion of 25,000 accepted steps or 25,000 rejected steps, whichever came first. The rmsd conformational clustering tolerance was 0.5 Å, calculated from the ligand's crystallographic coordinates. Standard deviations given in parentheses.

ΔG_{inter} the intramolecular energy, ΔG_{intra}, and the torsional free energy, ΔG_{tor}. These results are discussed case-by-case in what follows; "crystallographic rmsd" refers to the root-mean-square positional deviation of a given conformation from the crystallographic coordinates.

β-Trypsin / Benzamidine (3ptb)

The recognition of benzamidine by β-trypsin, which binds tightly in the specificity pocket of trypsin, is chiefly due to the polar amidine moiety

and the hydrophobic benzyl ring.[69] The amidine moiety was treated as being protonated. It was assumed that delocalization of the π-electrons of the benzene ring extended to the π-system of the amidine, and thus the ligand was treated as a rigid body. All three search methods succeeded in finding lowest energy conformations that were also the ones with the lowest crystallographic rmsd. In this case, the method that found the docked structure with the lowest energy was GA (Table V), but that found by the LGA method (Table VI) was practically the same. The mean of the final docked

TABLE V.
Results of Genetic Algorithm Dockings.[a]

PDB code	Number of clusters	Number in rank 1	Energy (kcal mol^{-1}) and rmsd (Å)						Number of energy evaluations
			Lowest energy	rmsd of lowest energy	Mean energy		Mean rmsd		
3ptb	2	9	−8.17	0.32	−7.72	(1.35)	1.50	(3.39)	1.35×10^6
2cpp	4	7	−7.36	0.93	−6.65	(2.11)	2.18	(3.42)	1.35×10^6
2mcp	10	1	−5.17	0.85	−3.61	(0.95)	5.26	(2.98)	1.35×10^6
1stp	7	4	−10.09	0.75	−8.42	(1.82)	2.96	(3.04)	1.35×10^6
1hvr	7	4	−21.41	0.82	−11.09	(9.79)	2.79	(1.97)	1.35×10^6
4hmg	9	2	−7.60	1.11	−5.72	(1.77)	2.32	(1.43)	1.35×10^6
4dfr	10	1	−16.10	0.95	−10.24	(3.95)	4.39	(2.37)	1.35×10^6

[a] The parameters used were 10 runs a population size of 50, and a run-termination criterion of a maximum of 27,000 generations or a maximum of 1.5 × 10⁶ energy evaluations, whichever came first. Note that, in this case all runs terminated after the maximum number of generations was reached, which equals the product of the population size and the number of generations. The rmsd conformational clustering tolerance was 0.5 Å, calculated from the ligand's crystallographic coordinates. Standard deviations are given in parentheses.

TABLE VI.
Results of Lamarckian Genetic Algorithm Dockings.[a]

PDB code	Number of clusters	Number in rank 1	Energy (kcal mol^{-1}) and rmsd (Å)						Number of energy evaluations
			Lowest energy	rmsd of lowest energy	Mean energy		Mean rmsd		
3ptb	1	10	−8.15	0.45	−8.15	(0.00)	0.46	(0.01)	1.50 × 10^6
2cpp	1	10	−7.36	0.93	−7.36	(0.00)	0.93	(0.00)	1.50 × 10^6
2mcp	6	2	−5.54	1.05	−4.15	(0.15)	1.10	(0.07)	1.50 × 10^6
1stp	1	10	−10.14	0.69	−10.08	(0.05)	0.66	(0.06)	1.50 × 10^6
1hvr	2	9	−21.38	0.76	−19.11	(6.92)	0.85	(0.35)	1.50 × 10^6
4hmg	3	7	−7.72	1.14	−7.54	(0.19)	1.18	(0.12)	1.50 × 10^6
4dfr	2	7	−16.98	1.03	−16.90	(0.07)	0.98	(0.07)	1.56 × 10^6

[a] The parameters used were 10 runs, a population size of 50, and a run-termination criterion of a maximum of 27,000 generations or a maximum of 1.5 × 10^6 energy evaluations, whichever came first. Because local search also uses energy evaluations, the total number of energy evaluations for the LGA method was greater than that for the GA method, using the same population size and maximum number of generations; in the LGA dockings, the runs terminated because the maximum number of energy evaluations was exceeded. The rmsd conformational clustering tolerance was 0.5 Å, calculated from the ligand's crystallographic coordinates. Standard deviations are given in parentheses.

energy across the ten dockings was lowest for LGA, followed by SA, and finally GA. This is reflected in a comparison of the mean rmsd of the docked conformation from the crystallographic structure for each of the methods: GA had the highest mean rmsd, followed by SA, and on average, the LGA produced conformations with the lowest crystallographic rmsd. Thus, considering their average performance, the best search method at finding the lowest energy and the lowest rmsd was the LGA. The predicted binding free energy, ΔG_{pred}, of the lowest docked energy structure obtained using the LGA method was −8.15 kcal mol^{-1} (Table VII), whereas the observed value, ΔG_{obs} was −6.46 kcal mol^{-1}; this is within the estimated error of the model.

Cytochrome P-450$_{cam}$ / Camphor (2cpp)

Camphor binds to the monooxygenase cytochrome P-450$_{cam}$ such that the 5-exo C—H bond is hydroxylated stereospecifically. The active site is deeply sequestered within the enzyme, and the crystal structure of the complex does not possess an obvious substrate access channel.[70] This buried active site presents a more challenging docking problem than 3ptb. Once bound, however, the substrate is "tethered" by a hydrogen bond that is donated from the Tyr-96 hydroxyl to the carbonyl oxygen of camphor, while the subtle complementarity of the pocket and the hydrophobic skeleton of camphor help to position the rest of the sub-

strate. The lowest energy found was −7.36 kcal mol^{-1}, found by both the GA (Table V) and LGA methods (Table VI); SA's lowest energy was −7.29 kcal mol^{-1} (Table IV), which is practically the same. All methods found the crystallographic structure, SA succeeding in 9 of 10 dockings, GA in 7 out of 10 dockings, and LGA in all of the dockings (with success, once again, being measured as having a crystallographic rmsd of less than 1 Å). In all three search method cases, the lowest energy cluster was the most populated, with 6, 9, and 10 members using SA, GA, and LGA, respectively. The predicted binding free energy, ΔG_{pred}, of the lowest docked energy structure, was −7.36 kcal mol^{-1} using the LGA method (see Table VII), whereas the observed value, ΔG_{obs} was −8.27 kcal mol^{-1}—once again, this was within the estimated error of the model.

McPC-603 / Phosphocholine (2mcp)

Antibody molecules bind their target antigens with exquisite specificity, having close complementarity between antigen and antibody surfaces, hydrogen bonding, van der Waals, and electrostatic interactions. Phosphocholine binds to Fab McPC-603,[71] and is an example of recognition is predominantly electrostatic in character, primarily due to the influence of Arg H52.[72] There is little conformational change in the side chains of Fab McPC-603 upon binding, as indicated from the unbound crystal structure. We allowed all four
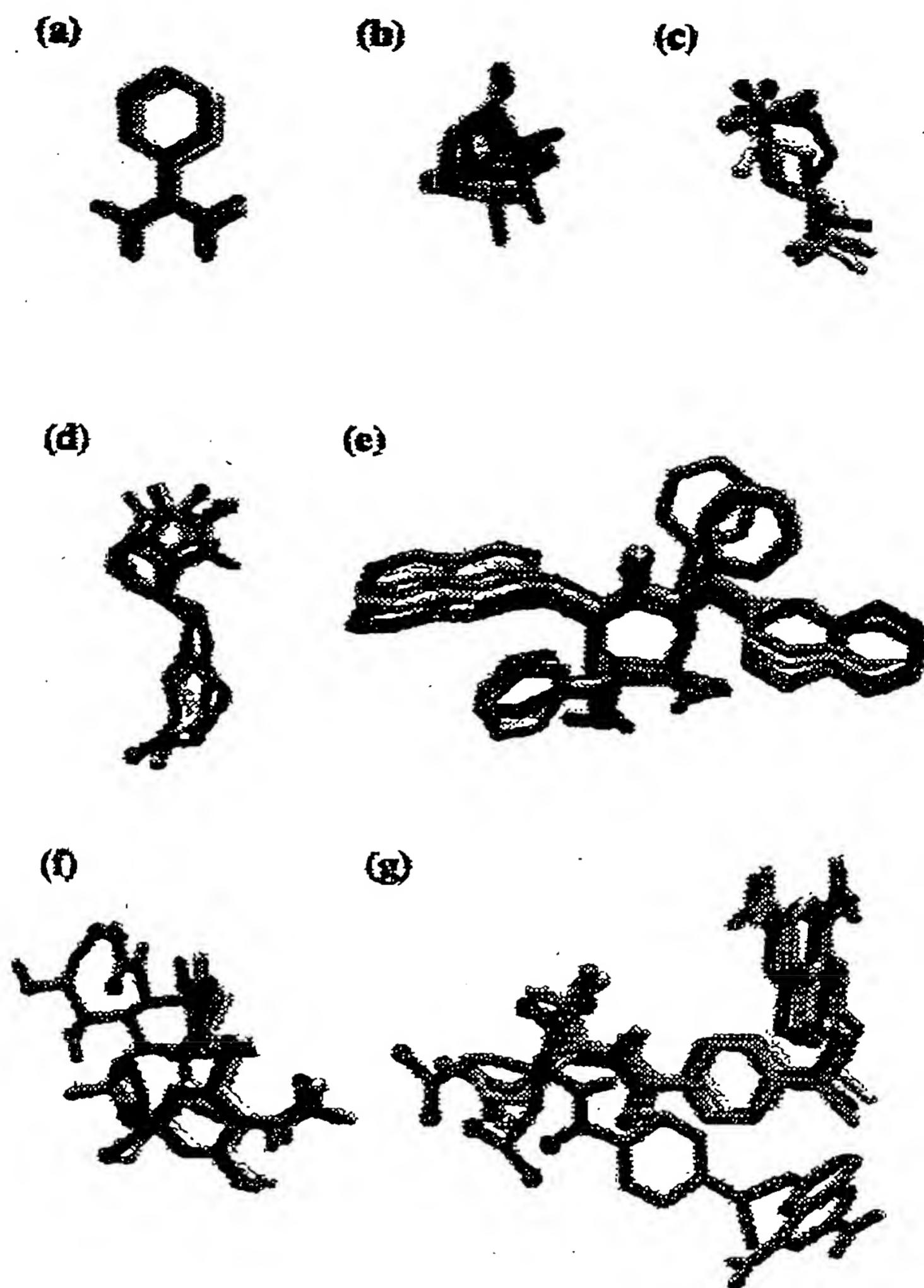
FIGURE 4. A comparison of the lowest energy structure found by each search method and the crystal structure. The latter is shown in black. The simulated annealing results are rendered with a striped texture, the genetic algorithm results are shaded gray, and the Lamarckian genetic algorithm results are white. Oxygen atoms are shown as spheres; other heteroatoms are not shown. Note that simulated annealing failed in the last two test cases, 4hmg and 4dfr, but both the genetic algorithm and the Lamarckian genetic algorithm succeeded.

bonds to rotate during docking. The energy of the crystal structure was positive, due in most part to a large, positive internal energy dominated by C2 and O1 being too close (2.26 Å); this could be improved if local minimization had been performed on the crystal structure before docking. The lowest energy found by each of the three search methods were −4.09, −5.17, and −5.54 kcal mol⁻¹, using SA, GA, and LGA, respectively. Unlike 3ptb and 2cpp, these differences in energy

were more significant. Both SA and GA found 10 different clusters, whereas LGA found 6 clusters. The mean energy of the 10 dockings was +70.89, −3.61, and −4.15 kcal mol⁻¹ for SA, GA, and LGA, respectively. Thus, on average, the LGA performed best in finding the lowest energy docked structure. Furthermore, the mean rmsd from the crystallographic coordinates was 5.40, 5.26, and 1.10 Å for SA, GA, and LGA, respectively, indicating that LGA also reproduced the crystal structure

TABLE VII.

Comparison of Predicted Free Energy of Binding, ΔG_{pred}, of Lowest Energy Docked Structure Obtained Using Lamarckian Genetic Algorithm, and Observed Free Energy of Binding, ΔG_{obs}.[a]

PDB code	Lowest energy	rmsd of lowest energy	ΔG_{inter}	ΔG_{intra}	ΔG_{tor}	ΔG_{pred}	ΔG_{obs}	$(\Delta G_{pred} - \Delta G_{obs})$
			Energy (kcal mol^{-1}) and rmsd (Å)					
3ptb	−8.15	0.45	−8.15	0.00	0.00	−8.15	−6.46	−1.69
2cpp	−7.36	0.83	−7.36	0.00	0.00	−7.36	−8.27	+0.91
2mcp	−5.54	1.05	−6.57	+1.03	+1.25	−5.32	−7.13	+1.81
1stp	−10.14	0.69	−9.90	−0.24	+1.56	−8.34	−18.27	+9.93[b]
1hvr	−21.38	0.76	−19.34	−2.04	+2.49	−16.85	−12.96	−3.89
4hmg	−7.72	1.14	−8.93	+1.21	+2.18	−6.75	−3.48	−3.27
4dfr	−16.97	1.03	−16.57	−0.40	+2.18	−14.39	−13.22	−1.17

[a] ΔG_{inter} is the intermolecular interaction energy between the ligand and the receptor, ΔG_{intra} is the intramolecular interaction energy of the ligand, and ΔG_{tor} is the torsional free energy change of the ligand upon binding.

[b] This large discrepancy may be due to neglect of the conformational rearrangements of streptavidin upon binding biotin, which are neglected in the docking simulation and binding free energy calculation.

most often. The predicted binding free energy, ΔG_{pred}, of the lowest docked energy structure was −5.32 kcal mol^{-1} using the LGA method (see Table VII), whereas the observed value, ΔG_{obs}, was −7.13 kcal mol^{-1}—this was also within the estimated error of the model.

Streptavidin / Biotin (1stp)

One of the most tightly binding noncovalent complexes is that of streptavidin/biotin, with an experimentally observed dissociation constant, K_d, of $10^{-15}M$. Comparison of the apo form and the complex[73] shows that the high affinity results from several factors, including formation of multiple hydrogen bonds and van der Waals interactions between the biotin and the protein, in addition to the ordering of surface polypeptide loops of streptavidin upon binding biotin. The method that found the lowest energy was LGA, at −10.14 kcal mol^{-1}, although GA was not significantly different, followed by SA with −8.48 kcal mol^{-1}. The method with the lowest mean energy was LGA at −10.06 kcal mol^{-1}, then GA with −8.42 kcal mol^{-1}, and finally SA with −7.76 kcal mol^{-1}. The method that found the crystallographic complex coordinates most often was LGA, having a mean rmsd of 0.66 Å, then SA at 1.24 Å, and finally GA with 2.96 Å. At the rmsd tolerance chosen for these experiments, 0.5 Å, SA found 10 different conformational clusters, GA found 7 clusters (the most populated was rank 1, with 4 members), and LGA found 1 cluster.

It was not possible to include the entropic effects of the flexible surface loops of streptavidin in the docking of biotin, although they make significant contributions to the binding free energy as revealed by a recent set of experiments involving an atomic force microscope.[74] It was found that the unbinding forces of discrete complexes of streptavidin with biotin analogs were proportional to the enthalpy change of the complex formation but independent of changes in the free energy, which indicates that the unbinding process is adiabatic and that entropic changes occur after unbinding. This may help to explain why the predicted binding free energy of the streptavidin/biotin complex (ΔG_{pred}) −10.14 kcal mol^{-1}, underestimated the magnitude of the observed value (ΔG_{obs}) −18.27 kcal mol^{-1} (Table VII).

HIV-1 Protease / XK263 (1hvr)

HIV-1 protease inhibitors prevent the maturation of virions of HIV, and are a major target for computer-assisted drug design in the development of AIDS therapies. Substrates and inhibitors of HIV-1 protease are typically extended peptides or peptidomimetics, with a dozen or more freely rotatable bonds and, as such, they present a challenging target for automated docking techniques. In addition, considerable protein motion is expected in the flaps upon binding, to allow the continuous polypeptide to reach the active site. However, most docking methods use a rigid protein target, and explicit modeling of the opening

and closing of the flaps is not performed: thus, the ligand must "thread" its way into the active site. The cyclic urea HIV-protease inhibitor, XK-263, has 10 rotatable bonds, excluding the cyclic urea's flexibility. All three search methods found solutions near to the crystal structure[75]: interestingly, the lowest docking energy found by SA was -11.77 kcal mol^{-1} and had an rmsd of 1.15 Å from the crystal structure, whereas GA and LGA found much lower energies but were still near to the active site, having crystallographic rmsd values of 0.82 Å and 0.76 Å, respectively. The lowest docking energy found overall was -21.41 kcal mol^{-1}, and was found using GA, although that found by LGA was practically the same. The predicted binding free energy, ΔG_{pred}, of the lowest energy structure was -16.85 kcal mol^{-1} using LGA, whereas the observed value, ΔG_{obs}, was -12.96 kcal mol^{-1}. The larger discrepancy between the predicted and observed values may be due to the entropic contributions of protein side chain and flap conformational rearrangements, or may be due to other low-energy conformational states of the cyclic urea moiety of XK-263, which are neglected in our calculations.

Influenza Hemagglutinin / Sialic Acid (4hmg)

The recognition of sialic acid by influenza hemagglutinin is chiefly mediated through hydrogen bonding: sialic acid has five hydroxyls, three in the glycerol group, one carboxylate, a cyclic ether oxygen, and an acetamido group, with a total of 11 rotatable acyclic bonds. We used the crystal structure of Weis et al.,[76] although the low resolution meant that the overall coordinate error was approximately 0.35–0.40 Å, which, in itself, presents a potential challenge in the docking tests. We modeled an isopropylated derivative of sialic acid to mimic part of an adjacent six-membered ring that would normally be present in this complex, but was not seen due to disorder; this introduced an extra rotatable bond, giving a total of 11 torsions. Furthermore, in these tests, we used the crystal conformation of the six-membered ring, although normally we would use several of the lowest energy conformations of the ring system and dock these separately.

This was one of two cases where simulated annealing failed to find a docking that was near the crystal structure: the lowest energy structure found had an rmsd of 3.77 Å from the crystallographic structure, and the docking with the lowest

crystallographic rmsd was 2.36 Å. The mean energy of all 10 SA dockings was very high (6.99 × 10^4 kcal mol^{-1}); only 4 of the 10 SA dockings found negative energies. Both GA and LGA, however, succeeded in finding conformations near ($\le$ 1.5 Å rmsd) the crystal conformation. The lowest energy found was by LGA, and was -7.72 kcal mol^{-1}. This structure had a crystallographic rmsd of 1.14 Å, and had a predicted binding free energy, ΔG_{pred}, of -6.75 kcal mol^{-1}; the observed binding free energy, ΔG_{obs}, for the sialic acid–hemagglutinin complex was -3.48 kcal mol^{-1}. The difference in predicted and observed binding free energies may be due to the structural differences between the isopropylated derivative that was docked and sialic acid itself.

Dihydrofolate Reductase / Methotrexate (4dgr)

Methotrexate is an antimetabolite that attacks proliferating tissue selectively induces remissions in certain acute leukemias[77]; however, dangerous side effects of methotrexate in normal cells continue to make DHFR an important target in computer-assisted anticancer drug design.[78] We used the crystal structure of E. coli dihydrofolate reductase complexed with methotrexate[79] to investigate a more challenging docking problem. We assumed that waters 603, 604, and 639, which mediate hydrogen bonding between the inhibitor and the protein, were conserved biowaters, and included them in the protein structure in our grid calculations. Ideally, these should be predicted, and recently a method based on a k-nearest-neighbors classifier and a genetic algorithm called Consolv was reported to do just this.[80]

This is one of the two test cases where simulated annealing failed: the lowest energy structure that it found had an rmsd of 4.83 Å from the crystal structure. This could be because the final docked conformation in simulated annealing is arrived at after a series of continuous steps, and if the route to the active site is blocked, the docking will tend to fail before the ligand reaches the active site. Note that, in the case of the camphor–cytochrome P-450$_{cam}$ docking, the random initialization loop was able to find initial states that were inside the binding pocket, but in this case the dockings failed to start near the active site.

The lowest energy found was -16.98 kcal mol^{-1}, and was found using LGA: this structure had an rmsd from the crystal structure of 1.03 Å.

The predicted binding free energy, ΔG_{pred}, of the lowest docked energy structure was -14.39 kcal mol^{-1} using LGA, whereas the observed value, ΔG_{obs}, was -13.22 kcal mol^{-1}. This finding was within the estimated error of the model.

JUDGING SEARCH METHODS

To evaluate the new search methods, and to compare them with the earlier search method of simulated annealing, we addressed the following questions: Which search method is most *efficient*? That is, which finds the lowest energy in a given number of energy evaluations? Which search method is most *reliable*? That is, which method finds the most conformations similar to that of the lowest energy? Finally, which search method is most *successful*? That is, which finds the crystallographic conformation most often after a given number of dockings? Furthermore, because these comparisons were carried out using the new, empirical free energy force field, these tests also represent an evaluation of the force field itself, and, if the global minimum of the force field is unable to reproduce observed crystallographic structures, its usefulness will be limited. Because it is very difficult to determine the global minimum of such a complex function, we cannot answer this question definitively; however, we can report the lowest energy found by any of the methods and its structural similarity to that of the crystal structure.

If we calculate statistics across all seven protein–ligand test systems for each search method, we obtain a quantitative estimate of relative performance of each search method (see Table VIII). If, in each of the seven test systems, we assume that the lowest docked energy found by any method is the effective global minimum energy, and then calculate the difference between this energy and all of the docked energies found by each search method, we can then calculate the mean and standard deviation of this difference energy for each search method. Ideally, the mean and standard deviation of this value would be zero. The mean of this difference energy was lowest for LGA (0.40 kcal mol^{-1}), followed by GA (3.41 kcal mol^{-1}), and finally SA (2.62 $\times$ 10^5 kcal mol^{-1}): the very high mean difference energy for SA is indicative of the cases in which this method failed to escape a local minimum, where the ligand was partially or wholly trapped within the protein. Hence, in answer to the first question, the Lamarckian genetic algorithm, LGA, is the most efficient search method.

In terms of how often the structure with lowest energy was found, LGA performed best: the mean of the number of docked structures in rank 1 was 78% for LGA, 40% for GA, and 24% for SA. The mean of the number of clusters found was lowest for LGA (2.29), followed by GA (7.00), and finally SA (8.43). Hence, the most reliable search method was LGA.

In comparing the relative success of each search method in reproducing the crystallographic structure, considering the crystallographic rmsd across all 10 dockings in each of the 7 test systems, the mean rmsd was lowest for LGA (0.88 Å, standard

TABLE VIII.
Statistical Comparison of Three Search Methods in AutoDock 3.0 Across all Seven Test Systems.[a]

Search method	Statistic	Number of clusters	Number in rank 1	Energy (kcal mol^{-1}) and rmsd (Å)				Number of energy evaluations
				Difference from effective global minimum energy	rmsd of lowest energy	Mean energy	Mean rmsd	
SA	Mean	8.43	2.43	2.62 $\times$ 10^5	1.85	2.61 $\times$ 10^5	3.63	1.75 $\times$ 10^6
	SD	2.70	2.44	1.40 $\times$ 10^5	1.74	4.60 $\times$ 10^4	2.61	4.15 $\times$ 10^5
GA	Mean	7.00	4.00	3.41	0.82	-7.64	3.06	1.35 $\times$ 10^6
	SD	3.06	3.06	5.31	0.25	2.59	1.32	0.00
LGA	Mean	2.29	7.86	0.40	0.86	-10.47	0.88	1.50 $\times$ 10^6
	SD	1.80	2.91	2.62	0.24	5.47	0.25	0.00

[a] The search methods are simulated annealing (SA), genetic algorithm (GA), and Lamarckian genetic algorithm (LGA). The mean and standard deviation (SD) for each criterion is shown. The effective global minimum energy for each of the seven test systems is the lowest docked energy found by any method for that test system. For each of the 10 dockings, the difference between the final docked energy and this effective global minimum energy was calculated; the mean and standard deviation was calculated across all 7 test systems, which was repeated for each search method.

TABLE IX
Results of Cross-Validation of Free Energy
Function Using Local Search on 20 HIV-1
Protease-Inhibitor Complexes.

PDB code	Experimental $\Delta G_{binding}$ (kcal / mol)	Calculated $\Delta G_{binding}$ (kcal / mol)
1hvs	−14.04	−10.95
1hvk	−13.79	−11.60
1hvi	−13.74	−12.39
7hvp	−13.11	−12.19
1hps	−12.57	−11.80
1hpv	−12.57	−8.24
4phv	−12.51	−14.36
1hef	−12.27	−9.52
1hiv	−12.27	−13.02
1hvl	−12.27	−10.35
8hvp	−12.27	−9.36
1aaq	−11.62	−9.68
1htg	−11.58	−13.13
9hvp	−11.38	−10.54
1hih	−10.97	−11.49
1heg	−10.56	−8.60
1sbg	−10.56	−10.35
1htf	−9.31	−8.21
1hbv	−8.68	−9.75
1hte	−7.69	−7.28

deviation 0.25 Å), followed by GA (3.06 Å, standard deviation 1.32 Å), and finally SA (3.63 Å, standard deviation 2.61 Å). These average results indicate that, of the three search methods, LGA will find the crystallographic structure most often. Thus, the answer to the last question, "Which method is most successful?," is LGA.

In two different cases, 4hmg and 4dfr, the simulated annealing method failed to reproduce the corresponding crystal structure, although it succeeded with 1hvr (see Fig. 4). This is important because methotrexate has 7 rotatable bonds, and would be expected to be solvable using our rule-of-thumb that SA succeeds in problems with 8 torsions or less; however, the HIV-1 protease inhibitor XK-263, has 10 rotatable bonds, and was successfully docked using SA. Thus, the degree of difficulty of a docking problem is not as simple as how many rotatable bonds there are; other factors, such as the nature of the energy landscape, clearly play an important role.

It could be said that the crystallographic rmsd of the lowest energy structure found by any of the search methods is an estimate of the quality of the force field, although this is complicated by the fact that the search method itself must determine a docking near to the global minimum, an unknown state. We can calculate the energy of the ligand in the crystal structure using the new force field (see Table II), which we assume to be near the global minimum, but, unfortunately, the crystal structure may contain frustrations and bad contacts. This appears to be the case in 2mcp, where a close contact between C2 and O1 causes a positive total energy to be calculated for the crystal structure. In all cases, the lowest energy found, considering all the search methods, was lower than that of the corresponding crystal structure.

The crystallographic rmsd of the lowest energy (found by any search method) for each of the protein–ligand test systems were all within 1.14 Å, or less, of the crystal structure. This suggests that the force field's global minimum in each of the protein–ligand cases was near to the crystal structure, if we accept the assumption that the crystal structure was near to or at the global minimum, and that the lowest energy found was near to the global minimum. In some cases, dockings were found that had lower crystallographic rmsd values but slightly higher energies than the lowest energy found. All of the lowest crystallographic rmsd values were 0.89 Å or less, indicating that low-energy structures found by the force field were very similar to the corresponding crystal structure.

Conclusion

AUTODOCK is a software package of general applicability for automated docking of small molecules, such as peptides, enzyme inhibitors, and drugs, to macromolecules, such as proteins, enzymes, antibodies, DNA, and RNA. New search methods have been introduced and tested here, using a new, empirical binding free energy function for calculating ligand–receptor binding affinities.

We have shown that, of the three search methods tested in AUTODOCK (simulated annealing, genetic algorithm, and Lamarckian genetic algorithm), the most efficient, reliable, and successful is the Lamarckian genetic algorithm LGA. We defined efficiency of search in terms of lowest energy found in a given number of energy evaluations; reliability in terms of reproducibility of finding the

lowest energy structure in independent dockings, as measured by the number of conformations in the top ranked cluster; and success in terms of reproducing the known crystal structure. Simulated annealing failed to reproduce the crystal structures for the influenza hemagglutin-sialic acid complex (4hmg) and the dihydrofolate reductase-methotrexate complex (4dfr). However, both the genetic algorithm and the Lamarckian genetic algorithm methods succeeded. Thus, the introduction of the LGA search method extends the power and applicability of AUTODOCK to docking problems with more degrees of freedom than could be handled by earlier versions.

The predicted binding affinities of the lowest energy docked conformations, using the LGA method and the new empirical free energy function, were within the standard residual error of the force field in four of the seven cases (3ptb, 2cpp, 2mcp, and 4dfr), and reasonably close in two other cases (1hvr and 4hmg). The large discrepancy between the predicted and the observed binding affinity of biotin for streptavidin (1stp), even though the crystal structure was successfully reproduced, may be due to the large free energy change that accompanies conformational changes in the protein upon binding, in particular the surface loops. This remains a limitation of the method, because protein motion is not modeled and successfully predicting such large-scale protein conformational changes is difficult. The AUTODOCK method works well when there is little change between the *apo* and ligand-bound forms of the protein, even if the protein undergoes significant conformational changes during binding.

AUTODOCK predicts the binding affinity using one conformation of the ligand-protein complex. A new class of models for predicting receptor-ligand binding affinities has been reported recently that considers not just the *lowest* energy state of the complex, but the *predominant states* of the binding molecules.[81] These approaches are grounded in statistical thermodynamics, and combine a modest set of degrees of freedom with aggressive conformational sampling to identify the low-energy conformations of the complex and the free molecules. AUTODOCK version 3.0 currently performs extensive conformational sampling, information that could be incorporated into the calculation of the binding affinity. We are studying how the search methods can be modified such that statistical thermodynamics calculations can be performed while the docking proceeds, to improve the calculation of the binding affinity.

Availability

Acknowledgments

References

1. J. M. Blaney, and J. S. Dixon, *Perspect. Drug Discov. Design*, 1, 301 (1993).

2. I. D. Kuntz, E. C. Meng, and B. K. Shoichet, *Acc. Chem. Res.*, 27, 117 (1994).

3. R. Rosenfeld, S. Vajda, and C. DeLisi, *Annu. Rev. Biophys. Biomol. Struct.*, 24, 677 (1995).

4. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, *J. Mol. Biol.*, 161, 269 (1982).

5. B. K. Shoichet and I. D. Kuntz, *Prot. Eng.*, 6, 723 (1993).

6. D. S. Goodsell and A. J. Olson, *Prot. Struct. Func. Genet.*, 8, 195 (1990).

7. G. M. Morris, D. S. Goodsell, R. Huey, and A. J. Olson, *J. Comput.-Aided Mol. Des.*, 10, 293 (1996).

8. N. Pattabiraman, M. Levitt, T. E. Ferrin, and R. Langridge, *J. Comput. Chem.*, 6, 432 (1985).

9. P. J. Goodford, *J. Med. Chem.*, 28, 849 (1985).

10. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.*, 21, 1087 (1953).

11. S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi, *Science*, 220, 671 (1983).

12. E. A. Lunney, S. E. Hagen, J. M. Domagala, C. Humblet, J. Kosinski, B. D. Tait, J. S. Warmus, M. Wilson, D. Ferguson, D. Hupe, P. J. Tummino, E. T. Baldwin, T. N. Bhat, B. Liu, and J. W. Erickson, *J. Med. Chem.*, 37, 2664 (1994).

13. J. V. N. Vara Prasad, K. S. Para, D. F. Ortwine, J. B. Dunbar Jr., D. Ferguson, P. J. Tummino, D. Hupe, B. D. Tait, J. M. Domagala, C. Humblet, T. N. Bhat, B. Liu, D. M. A. Guerin, E. T. Baldwin, J. W. Erickson, and T. K. Sawyer, *J. Am. Chem. Soc.*, 116, 6989 (1994).

14. A. R. Friedman, V. A. Roberts, and J. A. Tainer, *Prot. Struct. Func. Genet.*, 20, 15 (1994).

15. B. L. Stoddard and D. E. Koshland, *Nature*, 358, 774 (1992).

16. D. S. Goodsell, G. M. Morris, and A. J. Olson, *J. Mol. Recog.*, 9, 1 (1996).

17. C. M. Oshiro, I. D. Kuntz, and J. S. Dixon, *J. Comput.-Aided Mol. Design*, 9, 113 (1995).

18. D. R. Westhead, D. E. Clark, D. Frenkel, J. Li, C. W. Murray, B. Robson, and B. Waszkowycz, *J. Comput.-Aided Mol. Design*, 9, 139 (1995).

19. P. Willet, *TIBTECH*, 13, 516 (1995).

20. D. E. Clark and D. R. Westhead, *J. Comput.-Aided Mol. Design*, 10, 337 (1996).

21. C. D. Rosin, R. S. Halliday, W. E. Hart, and R. K. Belew, In *Proceedings of the Seventh International Conference on Genetic Algorithms (ICGA97)*, T. Baeck, Ed., Morgan Kauffman, San Francisco, CA, 1997.

22. D. K. Gehlhaar, G. M. Verkhivker, P. A. Rejto, C. J. Sherman, D. B. Fogel, L. J. Fogel, and S. T. Freer, *Chem. Biol.*, 2, 317 (1995).

23. J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, MI, 1975.

24. S. S. Četverikov, *J. Exper. Biol.*, 2, 3 (1926).

25. Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer-Verlag, New York, 1996.

26. W. E. Hart, *Adaptive Global Optimization with Local Search*, Ph.D. Thesis, Computer Science and Engineering Department, University of California, San Diego, 1994. See also: "ftp://ftp.cs.sandia.gov/pub/papers/wehart/thesis.ps.gz"

27. W. E. Hart, T. E. Kammeyer, and R. K. Belew, In *Foundations of Genetic Algorithms III*, D. Whitley and M. Vose, Eds., Morgan Kauffman, San Francisco, CA, 1994.

28. R. K. Belew and M. Mitchell, *Adaptive Individuals in Evolving Populations: Models and Algorithms. Santa Fe Institute Studies in the Science of Complexity*, XXVI, Addison-Wesley, Reading, MA, 1996.

29. F. J. Solis and R. J.-B. Wets, *Math. Oper. Res.*, 6, 19 (1981).

30. P.-G. Maillot, In *Graphics Gems*, A. S. Glassner, Ed., Academic Press, London, 1990, p. 498.

31. A. Watt and M. Watt, In *Advanced Animation and Rendering Techniques—Theory and Practice*, ACM Press, New York.

32. P. L'Ecuyer and S. Cote, *ACM Trans. Math. Software*, 17, 98 (1991).

33. J. B. Lamarck, *Zoological Philosophy*, Macmillan, London, 1914.

34. S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta Jr., and P. Weiner, *J. Am. Chem. Soc.*, 106, 765 (1984).

35. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *J. Am. Chem. Soc.*, 117, 5179 (1995).

36. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, *J. Comput. Chem.*, 4, 187 (1983).

37. A. T. Hagler, E. Huler, and S. Lifson, *J. Am. Chem. Soc.*, 96, 5319 (1977).

38. G. Némethy, M. S. Pottle, and H. A. Scheraga, *J. Phys. Chem.*, 87, 1883 (1983).

39. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. diNola, and J. R. Haak, *J. Chem. Phys.*, 81, 3684 (1984).

40. J. H. van der Waals, *Lehrbuch der Thermodynamik, Part 1*, Mass and Van Suchtelen, Leipzig, 1908.

41. I. K. McDonald and J. M. Thornton, *J. Mol. Biol.*, 238, 777 (1994).

42. M. K. Gilson and B. Honig, *Nature*, 330, 84 (1987).

43. D. Bashford and M. Karplus, *Biochemistry*, 29, 10219 (1990).

44. D. Bashford and K. Gerwert, *J. Mol. Biol.*, 224, 473 (1992).

45. B. Honig and A. Nicholls, *Science*, 268, 1144 (1995).

46. P. A. Bash, U. C. Singh, F. K. Brown, R. Langridge, and P. A. Kollman, *Science*, 235, 574 (1987).

47. L. P. Hammett, *J. Am. Chem. Soc.*, 59, 96 (1937).

48. T. Fujita, J. Iwasa, and C. Hansch, *J. Am. Chem. Soc.*, 86, 5175 (1964).

49. C. Hansch, A. R. Steward, J. Iwasa, and E. W. Deutsch, *Mol. Pharmacol.*, 1, 205 (1965).

50. C. D. Selassie, Z. X. Fang, R. L. Li, C. Hansch, G. Debnath, T. E. Klein, R. Langridge, and B. T. Kaufman, *J. Med. Chem.*, 32, 1895 (1989).

51. D. H. Williams, J. P. L. Cox, A. J. Doig, M. Gardner, U. Gerhard, P. T. Kaye, A. R. Lal, I. A. Nicholls, C. J. Salter, and R. C. Mitchell, *J. Am. Chem. Soc.*, 113, 7020 (1991).

52. L. Wesson and D. Eisenberg, *Prot. Sci.*, 1, 227 (1992).

53. H.-J. Böhm, *J. Comput.-Aided Mol. Design*, 6, 593 (1992).

54. H.-J. Böhm, *J. Comput.-Aided Mol. Design*, 8, 243 (1994).

55. A. N. Jain, *J. Comput.-Aided Mol. Design*, 10, 427 (1996).

56. E. L. Mehler and T. Solmajer, *Prot. Eng.*, 4, 903 (1991).

57. M. F. Sanner, A. J. Olson, and J.-C. Spehner, *Biopolymers*, 38, 305 (1996).

58. P. F. W. Stouten, C. Frommel, H. Nakamura, and C. Sander, *Mol. Simul.*, 10, 97 (1993).

59. S. Hill, In *Graphics Gems IV*, P. S. Heckbert, Ed., Academic Press, London, 1994, p. 521.

60. P. W. Atkins, *Physical Chemistry*, Oxford University Press, Oxford, 1982, p. 263.

61. H. R. Horton, L. A. Moran, R. S. Ochs, J. D. Rawn, and K. G. Scrimgeour, *Principles of Biochemistry*, Prentice-Hall, London, 1993.

62. S-PLUS, Statistical Sciences, Inc., Seattle, WA.

63. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, 112, 535 (1977).

64. E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, In *Crystallographic Databases-Information Content, Software Systems, Scientific Applications*, F. H. Allen, G. Bergerhoff, and R. Sievers, Eds., Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987, p. 107.

65. SYBYL, Tripos Associates, Inc., St. Louis, MO.

66. M. Marsili and J. Gasteiger, *Chim. Acta*, 52, 601 (1980).

67. J. Gasteiger and M. Marsili, *Tetrahedron*, 36, 3210 (1980).

68. D. N. A. Boobbyer, P. J. Goodford, P. M. McWhinnie, and R. C. Wade, *J. Med. Chem.*, 32, 1083 (1989).

69. M. Marquart, J. Walter, J. Deisenhofer, W. Bode, and R. Huber, *Acta Crystallogr. (Sect. B)*, 39, 480 (1983).

70. T. L. Poulos, B. C. Finzel, and A. J. Howard, *J. Mol. Biol.*, 195, 687 (1987).

71. E. A. Padlan, G. H. Cohen, and D. R. Davies, *Ann. Immunol. (Paris) (Sect. C)*, 136, 271 (1985).

72. J. Novotny, R. E. Bruccoleri, and F. A. Saul, *Biochemistry*, **28**, 4735 (1989).

73. P. C. Weber, D. H. Ohlendorf, J. J. Wendolski, and F. R. Salemme, *Science*, **243**, 85 (1989).

74. V. T. Moy, E.-L. Florin, and H. E. Gaub, *Science*, **266**, 257 (1994).

75. P. Y. S. Lam, P. K. Jadhav, C. J. Eyerman, C. N. Hodge, Y. Ru, L. T. Bacheler, J. L. Meek, M. J. Otto, M. M. Rayner, Y. Wong, C.-H. Chang, P. C. Weber, D. A. Jackson, T. R. Sharpe, and S. Erickson-Viitanen, *Science*, **263**, 380 (1994).

76. W. I. Weis, A. T. Brünger, J. J. Skehel, and D. C. Wiley, *J. Mol. Biol.*, **212**, 737 (1990).

77. C. K. Matthews and K. E. van Holde, *Biochemistry*, Benjamin/Cummings, Redwood City, CA, 1990.

78. C. A. Reynolds, W. G. Richards, and P. J. Goodford, *Anti-Cancer Drug Des.*, **1**, 291 (1987).

79. J. T. Bolin, D. J. Filman, D. A. Matthews, R. C. Hamlin, and J. Kraut, *J. Biol. Chem.*, **257**, 13650 (1982).

80. M. L. Raymer, P. C. Sanschagrin, W. F. Punch, S. Venkataraman, E. D. Goodman, and L. A. Kuhn, *J. Mol. Biol.*, **265**, 445 (1997).

81. M. K. Gilson, J. A. Given, and M. S. Head, *Chem. Biol.*, **4**, 87 (1997).

```
                1                                                          50
      Hi_Imp    ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~MSLRIK
      E_Coli    ~~~~~~~~~~  ~~~~~~~~~~  ~~MQSVTLCI  MPRQYLLTTL  VEILPMLRIA
       Strep    ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~MSNWDTKFL
   B_Subtilis   ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~MWESKFS
    Mt_Impdh    ~~~~~~~~~M  SRGMSGLEDS  SDLVVSPYVR  MGGLTTDPVP  TGGDDPHKVA
    Borellia    ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~    MPNKIT
   S. aureus    ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~MWESKFA


      Human2    ~~~~~~~~~~  ~~~~~~~~~~  ~~MADYLISG  GTSYVPDDGL  TAQQLFNC..
      Mouse2    ~~~~~~~~~~  ~~~~~~~~~~  ~~MADYLISG  GTSYVPDDGL  TAQQLFNC..
     Droimpdh   MESTTKVKVN  GFVESTSSSA  APAIQTKSTT  GFDAELQDGL  SCKELFQN..
       Yeast    ~~~~~~~~~~  ~~~~~~~MAA  IRDYKTALDL  TKSLPRPDGL  SVQELMDSKI
   Leishmania   ~~~~~~~~~~  ~~~~~~~~~~  ~~~MATNNAN  YRIKTIKDGC  TAEELFR...
        Arab    ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~MSTLEDGF  PADKLF..AQ



                51                                                        100
      Hi_Imp    QEALTFDDVL  LVPAHSTVLP  NTANLSTQLT  KEIRLNIPML  SAAMDTVTET
      E_Coli    KEALTFDDVL  LVPAHSTVLP  NTADLSTQLT  KTIRLNIPML  SAAMDTVTEA
       Strep    KKGYTFDDVL  LIPAESHVLP  NEVDLKTKLA  DNLTLNIPII  TAAMDTVTGS
   B_Subtilis   KEGLTFDDVL  LVPAKSEVLP  HDVDLSVELT  KTLKLNIPVI  SAGMDTVTES
    Mt_Impdh    MLGLTFDDVL  LLPAASDVVP  ATADTSSQLT  KKIRLKVPLV  SSAMDTVTES
    Borellia    KEALTFDDVS  LIPRKSSVLP  SEVSLKTQLT  KNISLNIPFL  SSAMDTVTES
   S. aureus    KESLTFDDVL  LIPAQSDILP  KDVDLSVQLS  DKVKLNIPVI  SAGMDTVTES


      Human2    GDGLTYNDFL  ILPGYIDFTA  DQVDLTSALT  KKITLKTPLV  SSPMDTVTEA
      Mouse2    GDGLTYNDFL  ILPGYIDFTA  DQVDLTSALT  KKITLKTPLV  SSPMDTVTEA
     Droimpdh   GEGLTYNDFL  ILPGYIDFTA  EEVDLSSPLT  KSLTLRAPLV  SSPMDTVTES
       Yeast    RGGLAYNDFL  ILPGLVDFAS  SEVSLQTKLT  RNITLNIPLV  SSPMDTVTES
   Leishmania   GDGLTYNDFI  ILPGFIDFGA  ADVNISGQFT  KRIRLHIPIV  SSPMDTITEN
        Arab    GYSYTYDDVI  FLPHFIDFST  DAVSLSTRLS  RRVPLSIPCV  SSPMDTVSES



                101                                                       150
      Hi_Imp    KLAISLAQEG  GIGFIHKNMT  IERQADRVRK  VKKFESGIVS  EPVTVLPNLT
      E_Coli    RLAIALAQEG  GIGFIHKNMS  IERQAEEVRR  VKKHESGVVT  DPQTVLPTTT
       Strep    KMAIAIARAG  GLGVIHKNMS  ITEQAEEVRK  VKRSENGVII  DPFFLTPEHK
   B_Subtilis   AMAIAMARQG  GLGIIHKNMS  IEQQAEQVDK  VKRSERGVIT  NPFFLTPDHQ
    Mt_Impdh    RMAIAMARAG  GMGVLHRNLP  VAEQAGQVEM  VKRSEAGMVT  DPVTCRPDNT
    Borellia    QMAIAIAKEG  GIGIIHKNMS  IEAQ......  ..........  ..........
   S. aureus    KMAIAMARQG  GLGVIHKNMG  VEEQADEVQK  VKRSENGVIS  NPFFLTPEES


      Human2    GMAIAMALTG  GIGFIHHNCT  PEFQANEVRK  VKKYEQGFIT  DPVVLSPKDR
      Mouse2    GMAIAMALTG  GIGFIHHNCT  PEFQANEVRK  VKKYEQGFIT  DPVVLSPKDR
     Droimpdh   EMAIAMALCG  GIGIIHHNCT  PEYQALEVHK  VKKYKHGFMR  DPSVMSPTNT
       Yeast    EMATFMALLD  GIGFIHHNCT  PEDQADMVRR  VKNYENGFIN  NPIVISPTTT
   Leishmania   EMAKTMALMG  GVGVLHNNCT  VERQVEMVKS  VKAYRNGFIS  KPKSVPPNTP
        Arab    HMAAAMASLG  GIGIVHYNCG  IAAQASIIRQ  AKSLKHPIAS  DAGVKFPEYE
```

Exhibit C

```
        151                                                     200
Hi_Imp      LAELAEMVKK  NGFAGYPVVD  ..G.ENNLIG  IITGRDTRFV  KDLSKTVSQV
E_Coli      LREVKELTER  NGFAGYPVVT  ..E.ENELVG  IITGRDVRFV  TDLNQPVSVY
Strep       VSEAEEELMQR  YRISGVPIVE  TLA.NRKLVG  IITNRDMRFI  SDYNAPISEH
B_Subtilis  VFDAEHLMGK  YRISGVPIVN  NEE.DQKLVG  IITNRDLRFI  SDYSMKISDV
Mt_Impdh    LAQVDALCAR  FRISGLPVVD  ..D.DGALVG  IITNRDMRFE  VDQSKQVAEV
Borellia    ...................................................
S. aureus   VYEAEALMGK  YRISGVPIVD  NKEDRNLVG  ILTNRDLRFI  EDFSIKIVDV

Human2      VRDVFEAKAR  HGFCGIPITD  TGRMGSRLVG  IISSRDIDFL  KEEEHDCFLE
Mouse2      VRDVFEAKAR  HGFCGIPITD  TGRMGSRLVG  IISSRDIDFL  KEEEHDRFLE
Droimpdh    VGDVLEARRK  NGFTGYPVTE  NGKLGGKLLG  MVTSRDIDF.  RENQPEVLLA
Yeast       VGEAKSMKEK  YGFAGFPVTA  DGKRNAKLVG  AITSRDIQFV  ..EDNSLLVQ
Leishmania  ISNIIRIKEE  KGISGILVTE  NGDPHGKLLG  IVCTKDIDYV  KNKDTP..VS
Arab        ITS.LDAFGP  SSFVFVEQTG  T.MTTPKLLG  YVTKSQWKRM  NYEQREMKIY


        201                                                     250
Hi_Imp      MTKKEDL  VTVKEGASRE  EILELMHQHR  VEKVLVVNDS  FKLKGMITVK
E_Coli      MTPKERL  VTVREGEARE  VVLAKMHEKR  VEKALVVDDE  FHLIGMITVK
Strep       MTSEHL..  VTAAVGTDLE  TAERILHEHR  IEKLPLVDNS  GRLSGLITIK
B_Subtilis  MTKEEL..  VTASVGTTLD  EAEKILQKHK  IEKLPLVDDQ  NKLKGLITIK
Mt_Impdh    MTKAPL..  ITAQEGVSAS  AALGLLRRNK  IEKLPVVDGR  GRLTGLITVK
Borellia    ...RKE..  IEKVKTYKFQ  KTINTNGDTN  EQKPEIFTAK  QHLEKSDAYK
S. aureus   MTQENL..  ITAPVNTTLE  EAEKILQKHK  IEKLPLVKD  GRLEGLITIK

Human2      EIMTKREDL.  VVAPAGITLK  EANEILQRSK  KGKLPIVNED  DELVAIIART
Mouse2      EIMTKREDL.  VVAPAGVTLK  EANEILQRSK  KGKLPIVNEN  DELVAIIART
Droimpdh    DIMTTEL.  VTAPNGINLP  TANAILEKSK  KGKLPIVNQA  GELVAMIART
Yeast       DVMTKNP...  VTGAQGITLS  EGNEILKKIK  KGRLLVVDEK  GNLVSMLSRT
Leishmania  AVMTRREKM.  TVERAPIQLE  EAMDVLNRSR  YGYLPIVNEN  DEVVNLCSRR
Arab        DYMKSCDSSD  YCVPWEIDFE  KLEFVLEDKQ  KGFV.VLERD  GETVNVVTKD


        251                                                     300
Hi_Imp      DFQKAEQKPN  ACKDEFG...  RLRVGAAVGA  GAGNEERIDA  LVKAGVDVLL
E_Coli      DFQKAEAKPN  ACKDEQG...  RLRVGAAVGA  GAGNEERVDA  LVAAGVDVLL
Strep       DIEKVIEFPH  AAKDEFG...  RLLVAAAVGV  TSDTFERAEA  LFEAGADAIV
B_Subtilis  DIEKVIEFPN  SSKDIHG...  RLIVGAAVGV  TGDTMTRVKK  LVEANVDVIV
Mt_Impdh    DFVKTEQHPL  ATKDSDG...  RLLVGAAVGV  GGDAWVRAMM  LVDAGVDVLV
Borellia    ..EHKEDFPN  ACKD..LN N  KLRVGAAVSI  DIDTIERVEE  LVKAHVDILV
S. aureus   DIEKVIEFPN  AAKDEHG...  RLLVAAAIGI  SKDTDIRAQK  LVEAGVDVLV

Human2      DLKKNRDYPL  ASKDA..K.K  QLLCGAAIGT  HEDDKYRLDL  LAQAGVDVVV
Mouse2      DLKKNRDYPL  ASKDA..K.K  QLLCGAAIGT  HEDDKYRLDL  LALAGVDVVV
Droimpdh    DLKKARSYPN  ASKDS..N.K  QLLVGAAIGT  RSEDKARLAL  LVANGVDVII
Yeast       DLMKNQKYPL  ASKSA..NTK  QLLWGASIGT  MDADKERLRL  LVKAGLDVVI
Leishmania  DAVRARDYPH  STLD...KSG  RLICAAATST  RPEDKRRVAA  LADVGVDVLV
Arab        DIQRVKGYPK  SGPGTVGPDG  EWMVGAAIGT  RESDKERLEH  LVNVGVNAVV
```

```
            301                                                              350
    Hi_Imp  IDSSHGHSEG VLQRVRETRA KY.PN.LPIV AGNVATAEGA IALADAGASA
    E_Coli  IDSSHGHSEG VLQRIRETRA KY.PD.LQII GGNVATAAGA RALAEAGCSA
     Strep  IDTAHGHSAG VLRKIAEIRA HF.PN.RTLI AGNIATAEGA RALYDAGVDV
 B_Subtilis IDTAHGHSQG VLNTVTKIRE TY.PE.LNII AGNVATAEAT RALIEAGADV
   Mt_Impdh VDTAHAHNRL VLDMVGKLKS EV.GDRVEVV GGNVATRSAA AALVDAGADA
   Borellia IDSAHGHSTR IIELIKKIKT KY P NLDLI AGNIVTKEAA LDLISVGADC
   S. aureus IDTAHGHSKG VIDQVKHIKK TY PE ITLV AGNVATAEAT KDLFEAGADI

     Human2 LDSSQGNSIF QINMIKYIKD KYP..NLQVI GGNVVTAAQA KNLIDAGVDA
     Mouse2 LDSSQGNSIF QINMIKYIKE KYP..SLQVI GGNVVTAAQA KNLIDAGVDA
    Droimpdh LDSSQGNSVY QVEMIKYIKE TYP..ELQVI GGNVVTRAQA KNLIDAGVDG
      Yeast LDSSQGNSIF QLNMIKWIKE TFP..DLEII AGNVVTKEQA ANLIAAGADG
  Leishmania LDSSQGNTIY QIAFIKWVKS TYP..HLEVV AGNVVTQDQA KNLIDAGADG
       Arab LDSSQGNSIY QLEMIKYVKK TYP..ELDVI GGNVVTMYQA QNLIQAGVDG


            351                                                              400
    Hi_Imp  VKVGIGPGSI CTTRIVTGVG VPQITAIADA AAALKDRGIP VIADGGIRFS
    E_Coli  VKVGIGPGSI CTTRIVTGVG VPQITAVADA VEALEGTGIP VIADGGIRFS
     Strep  VKVGIGPGSI CTTRVVAGVG VPQVTAIYDA AAVAREYGKT IIADGGIKYS
 B_Subtilis VKVGIGPGSI CTTRVVAGVG VPQITAIYDC ATEARKHGKT IIADGGIKFS
   Mt_Impdh VKVGVGPGSI CTTRVVAGVG APQITAILEA VAACRPAGVP VIADGGLQYS
   Borellia LKVGIGPGSI CTTRIVAGVG VPQITAICDV YEACNNTNIC IIADGGIRFS
   S. aureus VKVGIGPGSI CTTRVVAGVG VPQITAIYDC ATEARKHGKA IIADGGIKFS

     Human2 LRVGMGSGSI CITQEVLACG RPQATAVYKV SEYARRFGVP VIADGGIQNV
     Mouse2 LRVGMGSGSI CITQEVLACG RPQATAVYKV SEYARRFGVP VIADGGIQNV
    Droimpdh LRVGMGSGSI CITQEVMACG CPQATAVYQV STYARQFGVP VIADGGIQSI
      Yeast LRIGMGTGSI CITQKVMACG RPQGTAVYNV CEFANQFGVP CMADGGVQNI
  Leishmania IRIGMGSGSI CITQEVLACG RPQGTAVYKV AQYCASRGVP CTADGGLRQV
       Arab LRVGMGSGSI CTTQEVCAVG RGQATAVYKV CSIAAQSGIP VIADGGISNS


            401                                                              450
    Hi_Imp  GDIA.KAIAA GASCVMVGSM FAGTEEAPGE IELYQGRAFK SYRGMGSLGA
    E_Coli  GDIA.KAIAA GASAVMVGSM LAGTEESPGE IELYQGRSYK SYRGMGSLGA
     Strep  GDIV.KALAA GGNAVMLGSM FAGTDEAPGE TEIYQGRKFK TYRGMGSIAA
 B_Subtilis GDIT.KALAA GGHAVMLGSL LAGTSESPGE TEIYQGRRFK VYRGMGSVAA
   Mt_Impdh GDIA.KALAA GASTAMLGSL LAGTAEAPGE LIFVNGKQYK SYRGMGSLGA
   Borellia GDVV KAIAA GADSVMIGNL FAGTKESPSE EIIYNGKKFK SYVGMGSISA
   S. aureus GDII KALAA GGHAVMLGSL LAGTEESPGA TEIFQGRQYK VYRGMGSLGA

     Human2 GHIA.KALAL GASTVMMGSL LAATTEAPGE YFFSDGIRLK KYRGMGSLDA
     Mouse2 GHIA.KALAL GASTVMMGSL LAATTEAPGE YFFSDGIRLK KYRGMGSLDA
    Droimpdh GHIV.KAIAL GASAVMMGSL LAGTSEAPGE YFFSDGVRLK KYRGMGSLEA
      Yeast GHIITKALAL GSSTVMMGGM LAGTTESPGE YLYQDGKRLK AYRGMGSIDA
  Leishmania GDI.CKALAI GANCAMLGGM LSGTTETPGE YFFKGGVRLK VYRGMGSLEA
       Arab GHIV.KALVL GASTVMMGSF LAGSTEAPGG YEYTNGKRIK KYRGMGSLEA
```

```
              451                                                    500
    Hi_Imp    MA......KG  SSDRYFQ.SD  NAADKLVPEG  IEGRIPYKGY  LKEIIHQQMG
    E_Coli    MS......KG  SSDRYFQ.SD  NAADKLVPEG  IEGRVAYKGR  LKEIIHQQMG
     Strep    MK......KG  SSDRYFQGSV  NEANKLVPEG  IEGRVAYKGA  ASDIVFQMLG
 B_Subtilis   ME......KG  SKDRYFQ...  EENKKFVPEG  IEGRTPYKGP  VEETVYQLVG
   Mt_Impdh   MRGRGGATSY  SKDRYFADDA  LSEDKLVPEG  IEGRVPFRGP  LSSVIHQLTG
  Borellia    MK......RG  SKSRYFQLEN  NEPKKLVPEG  IEGMVPYSGK  LKDILTQLKG
  S. aureus   ME......KG  SNDRYFQ ED  KAPKKFVPEG  IEGRTAYKGA  LQDTIYQLMG

    Human2    MD....KHLS  SQNRYFSEAD  KIK...VAQG  VSGAVQDKGS  IHKFVPYLIA
    Mouse2    MD....KHLS  SQNRYFSEAD  KIK...VAQG  VSGAVQDKGS  IHKFVPYLIA
   Droimpdh   MERGDAKGAA  MSRYYHNEMD  KMK...VAQG  VSGSIVDKGS  VLRYLPYLEC
     Yeast    MQKTGTKGNA  STSRYFSESD  SVL...VAQG  VSGAVVDKGS  IKKFIPYLYN
 Leishmania   M....SQGKE  SGKRYLSENE  AVQ...VAQG  VSGNVVDKGS  AAKLIAYVSK
      Arab    MTKG......  SDQRYLGDQT  KLK...IAQG  VVGAVADKGS  VLKLIPYTMH


              501                                                    550
    Hi_Imp    GLRSCMGLTG  CATIDELRT.  ....KAEFVR  ISGAGIKESH  VHDVAITKEA
    E_Coli    GLRSCMGLTG  CGTIDELRT.  ....KAEFVR  ISGAGIQESH  VHDVTITKES
     Strep    GIRSGMGYVG  AGDIQELHE.  ....NAQFVE  MSGAGLIESH  PHDVQITNEA
 B_Subtilis   GLRSGMGYCG  SKDLRALRE.  ....EAQFIR  MTGAGLRESH  PHDVQITVHR
   Mt_Impdh   GLRAAMGYTG  SPTIEVLQ..  ....QAQFVR  ITPAGLKESH  PHDVAMTVEA
  Borellia    GLMSGMGYLG  AATISDL...  ..KINSKFVK  ISHSSLKESH  PHDVFSIT
  S. aureus   GVRAGMGYTG  SHDLRELRE.  ....EAQFTR  MGPAGLAESH  PHNIQITKES

    Human2    GIQHSCQDIG  AKSLTQVRAM  MYSGELKFEK  RTSSAQVEGG  VHSLHSYEKR
    Mouse2    GIQHSCQDIG  AKSLTQVRAM  TYSGELKFEK  RTSSAQVEGG  VHSLHSYEKR
   Droimpdh   GLQHSCQDIG  ANSINKLRDM  IYNGQLRFMK  RTHSAQLEGN  VHGLFSYEKR
     Yeast    GLQHSCQDIG  CRSLTLLKEN  VQSGKVRFEF  ~~~~~~~~~~  ~~~~~~~~~~
 Leishmania   GLQQSAQDIG  EISFDAIREK  MYAGQVLFSR  RSPTAQGEGG  VHSLHSYEKK
      Arab    AVKQGFQDLG  ASSLQSAHGL  LRSNILRLEA  RTGAAQVEGG  VHGLVSYEKK


              551                       581
    Hi_Imp    PNYRMG~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~
    E_Coli    PNYRLGS~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~
     Strep    PNYSVH~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~
 B_Subtilis   NKALPGLFGS  HQKKTGFVYD  ECCQSGFFSS  D
   Mt_Impdh   PNYYA~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~
  Borellia
  S. aureus   PNYSF

    Human2    LF~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~
    Mouse2    LF~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~
   Droimpdh   LF~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~
     Yeast    ~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~
 Leishmania   LFAAKM~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~
      Arab    SF~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~
```

Title: Amino Acid Identity in the Binding Pocket of IMPDH Enzymes

Exhibit D

Organism	Identical residues/total loop residues										% Identity
	Loop 1	Loop 2	Loop 3	Loop 4	Loop 5	Loop 6	Loop 7	Loop 8	Loop 9	All loops	
E. coli	6/7	6/6	2/7	6/9	3/4	19/21	3/3	51/69	2/7	98/133	74
S. aureus	6/7	4/6	3/7	8/9	4/4	21/21	3/3	46/69	4/7	99/133	74
B.subtilus	6/7	6/6	6/7	8/9	3/4	21/21	3/3	46/69	4/7	99/133	74
H. Influenzae	6/7	5/6	3/7	6/9	3/4	19/21	3/3	52/67	2/7	97/133	73
M. tuberculosis	6/7	2/6	4/7	5/9	2/4	19/21	3/3	38/69	3/7	79/133	59
Borellia	6/7	5/6	3/7	7/9	1/4	20/21	3/3	43/69	4/7	92/133	69
Human	4/7	2/6	2/7	3/9	2/4	12/21	3/3	28/69	2/7	56/133	42
Arabidopsis	4/7	2/6	1/7	3/9	1/4	12/21	3/3	27/69	2/7	53/133	40
Drosophila	4/7	2/6	1/7	3/9	1/4	12/21	3/3	26/69	2/7	52/133	39
S. cerevesia	4/7	2/6	1/7	3/9	1/4	11/21	3/3	29/69	1/7	54/133	41
L.donovani	3/7	2/6	0/7	2/9	1/4	11/21	3/3	27/69	1/7	49/133	37